Three Aspects of Biostatistical Learning Theory

Three Aspects of Biostatistical Learning Theory

A DISSERTATION PRESENTED
BY
MATEY NEYKOV NEYKOV
TO
THE DEPARTMENT OF BIOSTATISTICS

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY
IN THE SUBJECT OF
BIOSTATISTICS

HARVARD UNIVERSITY
CAMBRIDGE, MASSACHUSETTS
MAY 2015

Thesis advisors: Professors Jun S. Liu and Tianxi Cai Matey Neykov Neykov

Three Aspects of Biostatistical Learning Theory

Abstract

In the present dissertation we consider three classical problems in biostatistics and statistical learning — classification, variable selection and statistical inference.

Chapter 2 is dedicated to multi-class classification. We characterize a class of loss functions which we deem *relaxed Fisher consistent*, whose local minimizers not only recover the *Bayes rule* but also the exact conditional class probabilities. Our class encompasses previously studied classes of loss-functions, and includes non-convex functions, which are known to be less susceptible to outliers. We propose a generic greedy functional gradient-descent minimization algorithm for boosting *weak learners*, which works with any loss function in our class. We show that the boosting algorithm achieves geometric rate of convergence in the case of a convex loss. In addition we provide numerical studies and a real data example which serve to illustrate that the algorithm performs well in practice.

In Chapter 3, we provide insights on the behavior of *sliced inverse regression* in a high-dimensional setting under a *single index model*. We analyze two algorithms: a thresholding based algorithm known as *diagonal thresholding* and an L_1 penalization algorithm — *semidefinite programming*, and show that they achieve optimal (up to a constant) sample size in terms of support recovery in the case of standard Gaussian predictors. In addition, we look into the performance of the *linear regression LASSO* in single index models with correlated Gaussian designs. We show that under certain restrictions on the covariance and signal, the linear regression LASSO can also enjoy optimal sample size in terms of support recovery. Our analysis extends existing results on LASSO's variable selection capabilities for linear models.

Chapter 4 develops general inferential framework for testing and constructing confidence intervals for *high-dimensional estimating equations*. Such framework has a variety of applications and allows us to provide tests and confidence regions for parameters estimated by algorithms such as the *Dantzig Selector*, *CLIME* and *LDP* among others, non of which has been previously equipped with inferential procedures.

Contents

Listing of figures

To my parents — Sonya and Neyko.

Acknowledgments

I would like to extend many thanks to Professor Tianxi Cai, for she was always there to help in my continuous academic struggles, and was supportive in giving me the freedom to work on various exiting problems. Her help consisted not only of passionate discussions on statistics, but also of friendly chats and her marvelous yearly parties. During my time at Harvard I took the most classes with Professor Jun S. Liu as a lecturer. It is not that much of the class material*, than the philosophy of statistics that I learned from Jun, for which I will always be indebted to him. His critical thinking helped me develop the instincts of a researcher and I hope to carry the torch he lit for a long time. I am also truly thankful to Professor Xihong Lin, for patiently listening to all of my presentations and offering her insights, many of which were often opening entirely new horizons.

My discussions with Dr Qian Lin, were as fun as they were enlightening, and I owe him special thanks for his consistent help and encouragements. I was also very lucky to have the opportunity to collaborate with Professor Han Liu and Dr Yang Ning, so they have my full admirations. During my last semester I had several brief but extremely helpful chats with Professor Andrea Rotnitzky, which helped me understand the geometry of statistical inference, so thanks Andrea! I owe thanks to Professor Noureddine El Karoui, whom I randomly bumped into while purchasing double espressos, for providing me with his great insights into my problems. Thank you also to Rajarshi Mukherjee for organizing the student led seminars on statistics — the place to discuss cutting edge research and learn from your fellow students.

Next I want to mention several people, to whom I owe thanks from the bottom of my heart, because they directly or indirectly have motivated me, helped me and supported me throughout the years. I would like to thank Aleksandar Lishkov, for being a wonderful friend for almost my entire life. Thank you Lishke, without you going to Harvard would have been as impossible as surviving the years here. I also want to thank Rossen Kralev, a Harvard graduate himself, who hosted me in his NYC flat many times when I, for one reason or the other, were left homeless in Boston. Many fun summers would not have been the same if Ivaylo Boyadzhiev was not around the Boston area, so thanks Ivo! Kossyo Kokalanov, I thank you for bringing some amazing art in my dull academic life. Thank you Kiril Boyadzhiev for being an incredibly inspirational figure to me ever since 5th grade! Thank you Vlado Djonev, for keeping the good music flowing when I was working at the late hours of the Bulgarian night. My foosball skills were sharp due to years of extensive training along side with Petar Tashev, and enabled me and Caleb Miles to win the GSAS foosball tournament.

*Although, I definitely learned some of the class material as well ☺

Thank you Petar and Caleb! Ivan Topalov's sporadic visits to Boston, quickly turned into mini-adventures every time, so thanks Topalke for all the fun. Thank you Yered Pita-Juarez for driving me to IKEA, fixing my sink when it wouldn't stop running, moving me several times, driving me to IKEA again, but most of all I thank you for staying a true friend. I would also like to thank all students in the Biostatistics Department at Harvard, and in particular all students from my cohort, for asking hard questions in class, being easy-going out of class, and for contributing to the great and relaxed atmosphere in the Department. Thank you also goes to Seth Macfarlane, J. G. Quintel and Matt Groening for keeping my spirits up with their amusing TV shows.

This thesis would have been completely impossible if not for the help of the amazing Yuanyuan Shen. She was the one to get me out of my personal crisis, the one who was there for me at moments when everything seemed impossible. Thanks Ms Shen for the awesome dinners, but even better breakfasts. You showed me that traveling around the world can be a great way of letting statistics problems take a break from me, but even better — you taught me that living the PhD life can certainly be made a much more pleasurable experience, if we were to move Harvard to the Caribbean. Pura Vida!

Needless to say, I wouldn't have been in the position of writing these lines without the help and love from my family — Neyko, Sonya and Nadya Neykovi. It is due to them that I grew up to be naturally curious and was motivated to pursue the best education.

Last but not least, during my 5 years here, many people have pondered what exactly biostatistics is and why am I pursuing a PhD in this mysterious subject. While a definite answer would require a refinement of the question, one response that I have always been compelled to give is — "A career in biostatistics is completely justified in my case as my mom Sonya is a biologist and my dad Neyko is statistician." To you I dedicate this dissertation.

He who loves practice without theory is like the sailor who boards ship without a rudder and compass and never knows where he may cast.

Leonardo da Vinci

1 Introduction

Classification, variable selection and statistical inference are important areas in classical statistics. In this dissertation we place these fields in "modern" settings and provide some insights.

Chapter 2 is dedicated to classification. In particular, we focus on multi-class classification, which has been an active area of research. Accurate classification of categorical outcomes is essential in a wide range of applications. Due to computational issues with minimizing the 0/1 loss, Fisher consistent losses have been proposed as viable proxies. However, even with smooth losses, direct min-

imization remains a daunting task. To approximate such a minimizer, various boosting algorithms have been suggested. For example, with exponential loss, the AdaBoost algorithm [22] is widely used for two-class problems and has been extended to the multi-class setting [100]. Alternative loss functions, such as the logistic and the hinge losses, and their corresponding boosting algorithms have also been proposed [103,88]. In chapter 2, we demonstrate that a broad class of losses, including non-convex functions, can achieve Fisher consistency. While non-convex Fisher consistent losses have been previously discussed in the literature [6,76], the functions from our class possess the further property to recover the exact conditional probabilities. In addition, we provide a generic boosting algorithm that is not loss-specific. Having multiple boosting algorithms with different choices of loss functions, motivates a cross validation (CV) procedure to further improve the robustness of the proposed procedures. Simulation results suggest that the proposed boosting algorithms could outperform existing methods with properly chosen losses and the CV aggregation generally leads to classifiers with performances similar or better than any classifier with a pre-selected loss.

In Chapter 3 we look into high-dimensional variable selection in single index models, which we refer to as *support recovery*. Throughout the majority of the this chapter we explore support recovery algorithms based on Sliced Inverse Regression (SIR). SIR is a dimension reduction tool, which leverages information of an outcome variable, to project the predictor variables on a lower dimensional space containing all necessary information for prediction of the outcome. We study the behavior of SIR in a high-dimensional setting with $p \gg n$ under the assumption that the low dimensional space has dimension one. In particular, we provide two algorithms inspired by sparse principal component analysis (PCA) — diagonal thresholding (DT) and semidefinite programming (SDP), which achieve optimal sample size for support and signed support recovery correspondingly, up to a proportionality constant, under the assumption that the predictor matrix $X \sim N(0, \mathbb{I})$. In contrast, it is known that DT and SDP are sub-optimal in the PCA setting [3]. In addition, in chapter 3 we also explore two more algorithms – Covariance Thresholding (CT) and Linear LASSO's

performances in single index models. Under a slightly different set of assumptions to those in the SIR framework, we prove that CT can also provide signed support recovery in optimal (up to a constant) sample size. We also show that the Linear LASSO can recover the support of the single index model with a Gaussian predictor matrix $X \sim N(0, \Sigma)$, given that certain restrictions on Σ are met. The last complements existing results for the Linear LASSO support recovery[87], and can be viewed as a partial solution towards support recovery for predictor matrices with generic covariances.

In Chapter 4 we propose a novel inferential framework of testing hypotheses and constructing confidence regions for high-dimensional statistical models that can be fitted by solving a system of regularized estimating equations. Such an estimating equation based inferential framework is quite general and can be used for a wide variety of regularized estimators, including penalized M-estimators, constrained Z-estimators, and even greedy estimators. The key ingredient of this framework is a test statistic constructed by projecting the fitted estimating equations to a sparse direction obtained by solving a large-scale linear program. For hypothesis tests, we derive the limiting distribution of this proposed test statistic under both null and local alternative hypotheses. For confidence regions, we develop uniformly valid confidence intervals for low dimensional parameters of interest, and show their optimality under scenarios when the estimating equation is based on a log-likelihood function. To illustrate the usefulness of this framework, we further apply it to conduct inference for several constrained Z-estimators which have not been equipped with inferential power before, including the Dantzig selector for high-dimensional regression, the LDP estimator for high-dimensional discriminant analysis, the CLIME estimator for high-dimensional graphical models, and a regularized transition matrix estimator for high-dimensional vector autoregressive models. Compared with existing methods, our framework is the only one that is applicable for the latter three applications. We provide thorough numerical simulations and real data experiments to back up the developed theoretical results.

Inanimate objects can be classified scientifically into three major categories: those that don't work, those that break down and those that get lost.

Russel Baker

2

Fisher Consistency and Applications in a Multi-Class Setting

2.1 Introduction

Accurate classification of multi-class outcomes is essential in a wide range of applications. To construct an accurate classifier for the outcome $C \in \{1, ..., n\}$ based on a predictor vector $\boldsymbol{X}$, the

target is often to minimize a misclassification rate, which corresponds to a 0/1 loss. We assume that the data $(C, \boldsymbol{X}^T)^T$ is generated from a fixed but unknown distribution $\mathbb{P}$. Specifically, one would aim to identify $\mathbf{f} = \{f_1(\cdot), ..., f_n(\cdot)\}$ that maximizes the misclassification rate

$$\mathbb{L}(\mathbf{f}) = \mathbb{E}\left[I\{C \neq c_{\mathbf{f}}(\boldsymbol{X})\}\right] = \mathbb{P}\left\{C \neq c_{\mathbf{f}}(\boldsymbol{X})\right\}^{\star}, \tag{2.1.1}$$

under the constraint $\sum_j f_j(\boldsymbol{X}) = 0$, where $I(\cdot)$ is the indicator function and for any $\mathbf{f}$, $c_{\mathbf{f}}(\boldsymbol{X}) = \text{argmax}_j\, f_j(\boldsymbol{X})$. Obviously, $\mathbf{f}_{\text{Bayes}} = \{f_{\text{Bayes},j}(\cdot) = \mathbb{P}(C = j \mid \cdot) - n^{-1}, j = 1, ..., n\}$ minimizes (2.1.1). In practice, one may approximate the Bayes classifier $c_{\mathbf{f}_{\text{Bayes}}}(\cdot)$ by modeling $\mathbb{P}(C = j \mid \cdot)$ parametrically or non-parametrically. However, due to the curse of dimensionality and potential model mis-specification, such direct modeling may not work well when the underlying conditional risk functions are complex. On the other hand, due to discontinuity, direct minimization of the empirical 0/1 loss is often both computationally and statistically undesirable.

To overcome these challenges, many novel statistical procedures have been developed by replacing the 0/1 loss with a *Fisher consistent* loss ϕ such that its corresponding minimizer can be used to obtain the Bayes classifier. Lin [50] showed that a class of smooth convex functions can achieve Fisher consistency (FC) for binary classification problems. Zou et al. [104] further extended these results to the multi-class setting. Support vector machine methods have been shown to yield Fisher consistent results for both binary and multi-class settings [49,54]. Relying on these FC results, boosting algorithms for approximating the minimizers of the loss functions have also been proposed for specific choices of losses. Boosting algorithms search for the optimal solution by greedily aggregating a set of "weak-learners" $\mathcal{G}$ via minimization of an empirical risk, based on a loss function ϕ. The classical AdaBoost algorithm [22] for example is based on the minimization of the exponential loss, $\phi(x) = e^{-x}$, using the forward stagewise additive modeling (FSAM) approach. Hastie et al. [30]

[⋆] Here the expectation is taken with respect to unknown true distribution $\mathbb{P}$.

showed that the population minimizer of the AdaBoost algorithm corresponds to the Bayes rule $c_{\mathbf{f}_{\text{Bayes}}}(\cdot)$ for the two-class setting. Zhu et al.[100] extended this algorithm and developed the Stagewise Additive Modeling using a Multi-class Exponential (SAMME) algorithm for the multi-class case.

Most existing work on Fisher consistent losses focuses on convex functions such as $\phi(x) = e^{-x}$ and $\phi(x) = |1-x|_+$. However, there are important papers advocating the usage of non-convex loss functions, which we will briefly discuss here. In[15] inspired by Shen et al.[73] the authors explore SVM type of algorithms with the non-convex "ramp" loss instead of the typical "hinge" loss in order to speed up computations. In[6], the authors study the concept of "classification calibration" in the two-class case. Classification calibration of a loss can be understood as uniform Fisher consistency, along all possible conditional probabilities on the simplex. They demonstrate that non-convex losses such as $1 - \tan(kx), k > 0$ can be classification calibrated in the two class case. More generally, Tewari and Bartlett[76] extend classification calibration to the multiclass case, and provide elegant characterization theorems. We will draw a link between our work and the work of Tewari and Bartlett[76] in Section 2.2.

Asymptotically, procedures such as the AdaBoost based on these losses would lead to the optimal Bayes classifier, provided sufficiently large space of weak learner set $\mathcal{G}$. However, in finite samples, the estimated classifiers are often far from optimal, and the choice of the loss ϕ could greatly impact the accuracy of the resulting classifier. In this chapter, we consider a broad class of loss functions that are potentially non-convex and demonstrate that the minimizer of these losses can lead to the Bayes rules for multi-category classification, and in fact can be used to explicitly restore the conditional probabilities. Moreover, we define an iteration which leads to local minimizers of these non-convex losses, which as we argue, can also recover the Bayes rule. The last observation has important consequences in practice, as global minimization of non-convex losses remains a challenging problem. On the other hand, non-convex losses, although not commonly used in the existing literature, could be more robust to outliers[59]. The rest of the chapter is organized as follows. In section

2.2 we detail the conditions for the losses and their corresponding FC results. In settings where the cost of misclassification may not be exchangeable between classes, we generalize our FC results to a weighted loss that accounts for differential costs. In section 2.3, we propose a generic boosting algorithm for approximating the minimizers and study some of its numerical convergence aspects. Since the choices of ϕ would affect the classification accuracy in finite sample, we also propose a cross validation (CV) based procedure to construct an aggregated classifier to further improve the robustness of our procedures. In section 2.4 we present simulation results comparing the performance of our proposed procedures to that of some existing methods including the SAMME. We apply our proposed algorithms to identify subtypes of diabetic neuropathy with EMR data from the Partners Healthcare. These numerical studies suggest that our proposed methods, with properly chosen losses, could potentially provide more accurate classification. Additional discussions are given in section 2.5. Proofs of the theorems are provided in Appendix A.

2.2 Fisher Consistency for a general class of loss functions

In this section we characterize a broad class of loss functions which we deem relaxed Fisher consistent. This class encompasses previous classes of loss functions, provided in [104], but also admits non-convex loss functions.

2.2.1 Fisher Consistency for 0/1 Loss

Suppose the training data available consists of N realizations of $(C, \boldsymbol{X}^\intercal)^\intercal$, $\mathscr{D} = \{(C_i, \boldsymbol{X}_i^\intercal)^\intercal, i = 1, \ldots, N\}$. We assume that the data is drawn from a fixed, but unknown distribution $\mathbb{P}$, and all expectations throughout the chapter are taken with respect to $\mathbb{P}$. Moreover, we assume throughout

that:

$$\min_{j\in\{1,\ldots,n\}} \mathbb{P}(C=j|\boldsymbol{X})>0 : \mathbb{P} \text{ almost surely in } \boldsymbol{X}. \tag{2.2.1}$$

Assumption (2.2.1) states that there any class C has a chance to be drawn for all $\boldsymbol{X}$, except on a set of measure 0, where determinism in the class assignment is allowed. For a given C, define a corresponding $n \times 1$ vector $\boldsymbol{Y}_C = (I(C=1),...,I(C=n))^\intercal$. Under this notation, clearly $\boldsymbol{Y}_C^\intercal \mathbf{f}(\boldsymbol{X}) = f_C(\boldsymbol{X})$. For identifiability the following constraint is commonly used in the existing literature (e.g. see [45,104,100] among others):

$$\sum_{j=1}^{n} f_j(\cdot) = 0. \tag{2.2.2}$$

To identify optimal $\mathbf{f}(\cdot)$ to classify C based on $\mathbf{f}(\boldsymbol{X})$, we consider continuous loss functions ϕ as alternatives to the 0/1 loss and aim to minimize

$$\mathbb{L}_\phi(\mathbf{f}) = \mathbb{E}[\phi\{\boldsymbol{Y}_C^\intercal \mathbf{f}(\boldsymbol{X})\}] = \mathbb{E}[\phi\{f_C(\boldsymbol{X})\}] = \sum_{j=1}^{n} \mathbb{E}[\phi\{f_j(\boldsymbol{X})\}\mathbb{P}(C=j \mid \boldsymbol{X})], \tag{2.2.3}$$

under the constraint (2.2.2). The loss function ϕ is deemed *Fisher consistent* (FC) if the global minimizer (assuming it exists) $\mathbf{f}_\phi = \operatorname{argmin}_{\mathbf{f}:\sum_j f_j=0} \mathbb{L}_\phi(\mathbf{f})$ satisfies

$$c_{\mathbf{f}_\phi}(\boldsymbol{X}) = {}^{\dagger}c_{\mathbf{f}_{\text{Bayes}}}(\boldsymbol{X}). \tag{2.2.4}$$

Hence, with a FC loss ϕ, the resulting $\operatorname{argmax}_j \mathbf{f}_\phi(x)$ has the nice property of recovering the optimal Bayes classifier for the 0/1 loss. Clearly, the global minimizer $\mathbf{f}_\phi(x)$ also minimizes $\mathbb{E}[\phi\{f_C(\boldsymbol{X})\} \mid \boldsymbol{X}=x]$ for almost all x. With a given data $\mathscr{D}$, we may approximate $\mathbf{f}_\phi$ by minimizing the empirical

[†]Formally the"=" in (2.2.4) should be understood as "$\subseteq$". For the sake of simplicity, we keep this slight abuse of notation consistent throughout the chapter.

loss function

$$\widehat{L}_\phi(\mathbf{f}) = \frac{1}{N}\sum_{i=1}^{N}\phi\{\boldsymbol{Y}_{C_i}^{\intercal}\mathbf{f}(\boldsymbol{X}_i)\} = \frac{1}{N}\sum_{i=1}^{N}\phi(f_{C_i}(\boldsymbol{X}_i)) = \frac{1}{N}\sum_{j=1}^{n}\sum_{i=1}^{N}\phi(f_j(\boldsymbol{X}_i))\mathbb{I}(C_i = j),$$

to obtain $\widehat{\mathbf{f}} = \text{argmin}_{\mathbf{f}:\sum_j f_j=0}\,\widehat{L}_\phi(\mathbf{f})$.

Existing literature on the choice of ϕ focuses almost entirely on convex losses, important exceptions being [6,76]. Here, we propose a general class of ϕ to include non-convex losses and generalize the concept of FC as we defined in (2.2.4). Specifically, we consider all continuous ϕ satisfying the following properties:

$$\phi(x) - \phi(x') \geq (g(x) - g(x'))k(x') \quad \text{for all } x \in \mathbb{R}, x' \in S = \{z \in \mathbb{R} : k(z) \leq 0\}, \quad (2.2.5)$$

where g and k are both strictly increasing continuous functions, with $g(0) = 1, \inf_{x\in\mathbb{R}} g(x) = 0, \sup_{x\in\mathbb{R}} g(x) = +\infty, k(0) < 0$ and $\sup_{x\in\mathbb{R}} k(x) \geq 0$. This suggests[‡] that $\phi\{g^{-1}(\cdot)\}$ is continuously differentiable and convex on the set $g(S) = \{g(z) : z \in S\}$. However, ϕ itself is not required to be convex or differentiable. Extensively studied convex losses such as $\phi(x) = e^{-x}$ and $\phi(x) = \log(1 + e^{-x})$ both satisfy these conditions. For $\phi(x) = e^{-x}$, (2.2.5) would hold if we let $g(x) = e^x$ and $k(x) = -e^{-2x}$. For the logistic loss $\phi(x) = \log(1 + e^{-x})$, we may let $g(x) = e^{cx}$ and $k(x) = -\{ce^{cx}(1+e^x)\}^{-1}$ for any positive constant $c > 0$. Alternatively, $g(x) = e^x(1+e^x)/2$ and $k(x) = -2\{e^x(1+e^x)(1+2e^x)\}^{-1}$ would also satisfy (2.2.5) for the logistic loss. Our class of losses also allows non-convex functions. For example, $\phi(x) = \log(\log(e^{-x} + e))$ is a non-convex loss and (2.2.5) holds if $g(x) = e^x$ and $k(x) = -\{e^x(e^{x+1} + 1)\log(e^{-x} + e)\}^{-1}$. On an important note, we would like to mention that all three examples above can be seen to fall into the general class of classification calibrated loss functions in the two class case, as defined by [6] and

[‡]We provide a formal proof of this fact in Appendix A under Lemma A.0.1.

hence are FC in the two-class case. We will see a more general statement relating condition (2.2.5) to the two class classification calibration (see Remark 2.2.5 below).

Next, we extend the FC property (2.2.4), to allow for more generic classification rules. For a loss function ϕ, if there exists a functional $\mathcal{H}$ such that the minimizer of (2.2.3) has the property:

$$\operatorname*{argmax}_j \mathcal{H}\{f_{\phi,j}(\boldsymbol{X})\} = c_{\mathbf{f}_{\text{Bayes}}}(\boldsymbol{X}), \tag{2.2.6}$$

then we call it *relaxed* Fisher consistent (RFC). Obviously, the RFC property would still recover the Bayes classifier. Moreover FC losses are special cases of the RFC losses with an increasing $\mathcal{H}$.

We will now point out a connection between RFC and multiclass classification calibration as defined by Tewari and Bartlett [76]. Re-casting the definition of multiclass classification calibration to our framework, it requires that for any vector $\boldsymbol{w}$ on the simplex, the minimizer (assuming it exists):

$$\widehat{\boldsymbol{F}}(\boldsymbol{w}) = \operatorname*{argmin}_{\boldsymbol{F}:\sum_j F_j=0} \sum_{i=1}^{n} \phi(F_j) w_j \text{ satisfies } \operatorname*{argmax}_j \mathcal{H}(\phi(\widehat{F}_j)) = \operatorname*{argmax}_j w_j, \tag{2.2.7}$$

for some functional $\mathcal{H}$. In words, classification calibration ensures that regardless of the conditional distribution of $C|\boldsymbol{X}$, one can recover the Bayes rule. In contrast, RFC requires this to happen for the distribution at hand $C|\boldsymbol{X}$, for ($\mathbb{P}$ almost) all $\boldsymbol{X}$. This subtle but important distinction makes a difference. Example 4 in [76] shows that if ϕ is positive and convex the conditions in (A.0.4) cannot be met for all vectors $\boldsymbol{w}$ on the simplex, when we have at least 3 classes. On the contrary, in the present chapter we argue that in fact condition (A.0.4) remains plausible for both convex and non-convex losses, provided that we require the assumption that the points $\boldsymbol{w}$ are not allowed to be verteces of the simplex (i.e. $w_j > 0$ for all j), which relates back to assumption (2.2.1).

The next result justifies that the proposed losses satisfying (2.2.5) are RFC with $\mathcal{H}(x) = H_\phi(x) \equiv g(x)k(x)$. We first present in Theorem 2.2.1 the property of a general constrained minimization

problem, which is key to establishing the RFC.

Theorem 2.2.1. *For a loss ϕ satisfying (2.2.5), consider the optimization problem with some given $w_j > 0$:*

$$\min_{\boldsymbol{F}=(F_1,\ldots,F_n)^\intercal} \sum_{j=1}^{n} \phi(F_j) w_j \quad \textit{under the constraint} \quad \prod_{j=1}^{n} g(F_j) = 1. \tag{2.2.8}$$

Assume that there exists a minimum denoted by $\widehat{\boldsymbol{F}} = (\widehat{F}_1, \ldots, \widehat{F}_n)^\intercal$. Then the minimizer $\widehat{\boldsymbol{F}}$ must satisfy

$$H_\phi(\widehat{F}_j) w_j = \mathcal{C} \quad \textit{for some } \mathcal{C} < 0. \tag{2.2.9}$$

Moreover, if the function $H_\phi(\cdot)$ is strictly monotone there is a unique point with the property described above.

This result indicates that $H_\phi(\widehat{F}_j)$ is inversely proportional to the weight w_j. Now, consider $g(x) = \exp(x)$, $w_j = \mathbb{P}(C = j | \boldsymbol{X} = x)$, and $F_j = f_j(x)$, where we hold x fixed, as in the boosting framework. Then we can recover $c_{\mathbf{f}_{\text{Bayes}}}(x)$ by classifying C according to $\operatorname{argmax}_j \{-H_\phi(\widehat{F}_j)\}^{-1} = \operatorname{argmax}_j H_\phi(\widehat{F}_j)$ (the negative sign comes in because $\mathcal{C} < 0$), which implies that ϕ is RFC. Note that when $H_\phi(\cdot)$ is not increasing, Theorem 2.2.1 does not immediately imply that ϕ is a Fisher consistent loss according to definition (2.2.4), because the Bayes classifier need not be recovered by $\operatorname{argmax}_j \widehat{F}_j$. Nevertheless, we have the following:

Proposition 2.2.2. *Assume the same conditions as in Theorem 2.2.1. Then in addition to (2.2.9) we have:*

$$\operatorname*{argmax}_{j \in \{1,\ldots,n\}} \widehat{F}_j = \operatorname*{argmax}_{j \in \{1,\ldots,n\}} w_j,$$

and hence ϕ is also FC in the sense of (2.2.4).

The validity of Proposition 2.2.2 can be deduced from Theorem 2.2.1 and an application of Lemma 4 of[76], but for completeness we include a simple standalone proof in Appendix A. While Proposition 2.2.2 states that ϕ is FC, Theorem 2.2.1 suggests that one can additionally recover the exact conditional probabilities by calculating:

$$w_j = \frac{\{H_\phi(\widehat{F}_j)\}^{-1}}{\sum_{j=1}^{n}\{H_\phi(\widehat{F}_j)\}^{-1}}. \tag{2.2.10}$$

It is also worth noting here that the constraint in (2.2.8), generalizes the typical identifiability constraint (2.2.2), and the two coincide when $g(\cdot) = \exp(\cdot)$. We proceed by formulating a sufficient condition for the optimization problem in Theorem 2.2.1 to have a minimum without requiring the convexity or differentiability of ϕ.

Theorem 2.2.3. *The optimization problem in Theorem 2.2.1 has a minimum if either of the following conditions holds:*

i. ϕ *is decreasing on the whole* $\mathbb{R}$ *and for all* $c > 0$*:*

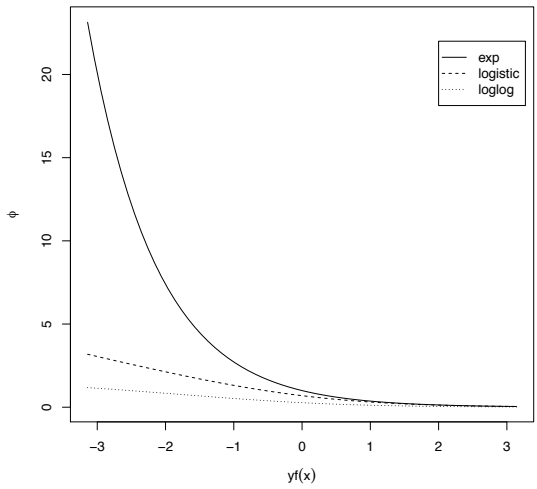
$$c\phi(g^{-1}(x)) + \phi(g^{-1}(x^{1-n})) \uparrow +\infty, \text{ as } x \downarrow 0, \tag{2.2.11}$$

ii. ϕ *is not decreasing on the whole* $\mathbb{R}$*.*

Remark 2.2.4. *It follows that in any case, problem (2.2.8) has a minimum when the loss function is bounded from below and unbounded from above.*

Remark 2.2.5. *Take* $g = \exp$ *to match the constraint in (2.2.8) with the constraint considered by Bartlett et al.*[6]*. It turns out that a loss function obeying (2.2.5) and either i. or ii. is classification calibrated in the two class case. See Lemma A.0.3 in Appendix A for a formal proof of this fact.*

Figure 2.1: Loss Functions Comparison



Clearly, by Remark 2.2.5, problem (2.2.8) would have a minimum for all three losses suggested earlier — the exponential, logistic (for both $g(x) = e^{cx}$, and $g(x) = e^x \frac{e^x+1}{2}$), and log-log loss.

Finally we conclude this subsection, by noting that the assumptions in both Theorem 1 and 2 in[104] can be seen to imply that the assumptions in Theorems 2.2.1 and 2.2.3 hold, thus rendering these theorems as consequences of the main result shown above. For completeness we briefly recall what these conditions are. In Theorem 1, Zou et al.[104] require a twice differentiable loss function ϕ such that $\phi'(0) < 0$ and $\phi'' > 0$. In Theorem 2 these conditions are slightly relaxed by allowing for part linear and part constant convex losses.

2.2.2 Fisher Consistency for Weighted 0/1 Loss

Although the expected 0/1 loss or equivalently the overall misclassification is an important summary for the overall performance of a classification, alternative measures may be preferred when the cost of misclassification is not exchangeable across outcome categories. For such settings, it would be

desirable to incorporate the differential cost when evaluating the classification performance and consider a weighted misclassification rate. Consider a cost matrix $\mathcal{W} = [W(j,\jmath)]_{n\times n}$ with $W(j,\jmath)$ representing the cost in classifying the $\jmath^{\text{th}}$ class to the j^{th} class. Then, the optimal Bayes classifier is

$$c^{\mathcal{W}}_{\mathbf{f}_{\text{Bayes}}}(\boldsymbol{X}) = \operatorname*{argmin}_{j} \sum_{\jmath=1}^{n} W(j,\jmath)\mathbb{P}(C=\jmath|\boldsymbol{X}). \tag{2.2.12}$$

Setting $\mathcal{W} = 1 - \mathcal{I}$ corresponds to the 0/1 loss and $c^{\mathcal{W}}_{\mathbf{f}_{\text{Bayes}}} = c_{\mathbf{f}_{\text{Bayes}}}$, where $\mathcal{I}$ is the identity matrix. Without loss of generality, we assume that $W(j,\jmath) \geq 0$. For ϕ satisfying (2.2.5) and the condition in Theorem 2.2.3, we next establish the FC results for the weighted 0/1 loss parallel to those given in Theorems 2.2.1 and 2.2.3.

Proposition 2.2.6. *Define the weighted loss $\ell(F_\jmath) = \sum_{j=1}^{n} \phi(F_j)W(j,\jmath)$. Then the optimization problem:*

$$\min_{\boldsymbol{F}=(F_1,\ldots,F_n)^\intercal : \prod_{j=1}^{n} g(F_j)=1} \sum_{\jmath=1}^{n} \ell(F_\jmath)\mathbb{P}(C=\jmath|\boldsymbol{X}), \tag{2.2.13}$$

has a minimizer $\widehat{\boldsymbol{F}}^{\mathcal{W}} = (\widehat{F}_1^{\mathcal{W}},\ldots,\widehat{F}_n^{\mathcal{W}})^\intercal$ which satisfies the property that:

$$H_\phi(\widehat{F}_j^{\mathcal{W}})w_j^{\mathcal{W}} = \widetilde{\mathcal{C}} \quad \textit{for some } \widetilde{\mathcal{C}} < 0, \tag{2.2.14}$$

where $w_j^{\mathcal{W}} = \sum_{\jmath=1}^{n} W(j,\jmath)\mathbb{P}(C=\jmath|\boldsymbol{X})$ assuming that $w_j^{\mathcal{W}} > 0$.

Proof of Proposition 2.2.6. This proposition is a direct consequence of Theorem 2.2.1, after exchanging summations:

$$\sum_{\jmath=1}^{n}\sum_{j=1}^{n} \phi(F_j)W(j,\jmath)\mathbb{P}(C=\jmath|\boldsymbol{X}) = \sum_{j=1}^{n} \phi(F_j)w_j^{\mathcal{W}}.$$

□

Remark 2.2.7. *Note that the above result hints on how one can relax assumption (2.2.1) by using the loss ℓ constructed with $\mathcal{W} = 1 - \mathcal{I}$. Using this particular ℓ, Proposition 2.2.6 simply requires $w_j^{\mathcal{W}} > 0$, which would be satisfied if we required:*

$$\max_{j\in\{1,\dots,n\}} \mathbb{P}(C = j|\boldsymbol{X}) < 1 : \mathbb{P} \textit{ almost surely in } \boldsymbol{X},$$

which is indeed weaker than (2.2.1). If we wanted to recover the probabilities simply note that $\mathbb{P}(C = j|\boldsymbol{X}) = 1 - w_j^{\mathcal{W}}$, and hence the probabilities can be recovered by:

$$\mathbb{P}(C = j|\boldsymbol{X}) = 1 - \frac{\{H_\phi(\widehat{F}_j^{\mathcal{W}})\}^{-1}}{\sum_{j=1}^n \{H_\phi(\widehat{F}_j^{\mathcal{W}})\}^{-1}}.$$

The result also suggests that using the modified loss ℓ, we can attain the optimal weighted Bayes classifier $c^{\mathcal{W}}_{\mathbf{r}_{\text{Bayes}}}(\boldsymbol{X})$ based on $\operatorname{argmin}_j H_\phi(\widehat{F}_j^{\mathcal{W}})$.

2.3 Generic Algorithm for Constructing the Classifer

In this section we provide a generic boosting algorithm, based on the explicit structure (2.2.5) that the RFC loss functions posses, and analyze certain numerical convergence aspects of the algorithm in the special case when $g = \exp$. We finish the section with a suggestion for aggregating boosted classifiers based on different loss functions.

2.3.1 A Generic Boosting Algorithm

The properties of ϕ and the results in Theorem 2.2.1 and 2.2.3 also lead to a natural iterative generic boosting algorithm to attain the minimizer.

A Conditional Iteration

In this subsection, we provide an iterative procedure, conditional on $\boldsymbol{X} = x$, which eventually leeds to a generic boosting algorithm. The usefulness of this conditional iteration is based on the following result.

Theorem 2.3.1. *Assume that ϕ satisfies (2.2.5) and the condition in Theorem 2.2.3. Starting from $\boldsymbol{F}^{(0)} \equiv 0$, i.e. $F_j^{(0)} = 0$ for all j, define the following iterative procedure:*

$$\boldsymbol{F}^{(m+1)} = \underset{\boldsymbol{F}:\prod g(F_j)=1}{\operatorname{argmax}} \sum_{j=1}^{n} \{g(F_j^{(m)}) - g(F_j)\} k(F_j) w_j. \tag{2.3.1}$$

This iteration is guaranteed to converge to a point $\boldsymbol{F}^*$ with the following property:

$$g(F_j^*)k(F_j^*)w_j = H_\phi(F_j^*)w_j = \mathcal{C} < 0. \tag{2.3.2}$$

Remark 2.3.2. *On an important note, careful inspection of the proof of Theorem 2.3.1, implies that in fact any iteration having $\boldsymbol{F}^{(m+1)}$ such that, $k(F_j^{(m+1)}) \in S$, $\prod_{j=1}^n g(F_j^{(m+1)}) = 1$, and:*

$$\sum_{j=1}^{n} \{g(F_j^{(m)}) - g(F_j^{(m+1)})\} k(F_j^{(m+1)}) w_j > 0,$$

will converge to a point with the property (2.3.2). This is important, as it implies that even if problem (2.3.1) is difficult to solve in practice, one can solve the simpler problem above, and will still arrive at a local minimum satisfying (2.3.2).

In the theorem above, the iterations are defined conditionally on $\boldsymbol{X} = x$, and F_j can be understood as $f_j(x)$. If $H_\phi(\cdot)$ turns out to be monotone, the procedure above will converge to the global minimum, as we can conclude straight from Theorem 2.2.1. Even if the procedure does not converge to a global minimum, because of the property of the point that it converges to, $\boldsymbol{F}^*$ can be used to

recover the Bayes classifier. This observation is particularly useful for minimizing non-convex loss functions as in such cases it is often hard to arrive at the global minimum. Moreover, as before the point $\boldsymbol{F}^*$ can be used not only for classification purposes, but also to recover the exact probabilities w_j.

In practice, the procedure described in (2.3.1) can be used to derive algorithms for boosting. However, an unconditional version of (2.3.1) is needed since w_j are unknown in general. Noting that the expectation of $I(C = j)$ given $\boldsymbol{X}$ is w_j, we have

$$\begin{aligned}
&\mathbb{E}\left(\sum_{j=1}^{n}\left[g\{F_j^{(m)}(\boldsymbol{X})\} - g\{F_j(\boldsymbol{X})\}\right]k\{F_j(\boldsymbol{X})\}w_j\right)\\
=&\mathbb{E}\left(\sum_{j=1}^{n}\left[g\{F_j^{(m)}(\boldsymbol{X})\} - g\{F_j(\boldsymbol{X})\}\right]k\{F_j(\boldsymbol{X})\}I(C=j)\right)\\
=&\mathbb{E}\left(\left[g\{\boldsymbol{Y}_C^\intercal\boldsymbol{F}^{(m)}(\boldsymbol{X})\} - g\{\boldsymbol{Y}_C^\intercal\boldsymbol{F}(\boldsymbol{X})\}\right]k\{\boldsymbol{Y}_C^\intercal\boldsymbol{F}(\boldsymbol{X})\}\right), \qquad (2.3.3)
\end{aligned}$$

where $\boldsymbol{Y}_C = (I(C = 1), ..., I(C = n))^\intercal$. We next derive a boosting algorithm iterating based on an empirical version of (2.3.3).

The Boosting Algorithm

To derive the boosting algorithm, we let $\mathcal{G} = \{G_b(\cdot), b = 1, ..., B\}$ denote the bag of weak learners with $G(\boldsymbol{X}) \in \{1, ..., n\}$ denoting the predicted class based on learner G. For the bth classifier in $\mathcal{G}$, define a corresponding vectorized version of G_b, $\boldsymbol{F}_b = (F_{b1}, ..., F_{bn})$, with

$$F_{bj}(\boldsymbol{X}) = \mathcal{C}_- + I\{G_b(\boldsymbol{X}) = j\}(\mathcal{C}_+ - \mathcal{C}_-),$$

where $\mathcal{C}_- < 0$ and $\mathcal{C}_+ > 0$ are chosen such that $\prod_{j=1}^{n} g(F_{bj}) = 1$. Obviously, $\boldsymbol{Y}_C^\intercal\boldsymbol{F}_b(\cdot) = \mathcal{C}_- + I(G_b(\cdot) = C)(\mathcal{C}_+ - \mathcal{C}_-)$. Let $\mathcal{G}^* = \{\boldsymbol{F}_b(\cdot), b = 1, ..., B\}$ denote the bag of vectorized

classification functions corresponding to the classifiers in $\mathcal{G}$.

We next propose a generic iterative boosting algorithm that greedily searches for an optimal weight and for which weak learner to aggregate at each iteration. The loss function ϕ is not directly used, and instead we rely on the $g(\cdot)$ and $k(\cdot)$ functions as specified in (2.2.5). Specifically, initialize $\boldsymbol{F}^{(0)} = 0$ and let $\mathcal{C}^{(0)} = 0$. Then for $m = 1, ..., M$ with M being the total number of desired iterations, we obtain the maximizer of

$$\sum_{i=1}^{N} g\{\boldsymbol{Y}_{Ci}^{\intercal}\boldsymbol{F}^{(m-1)}(\boldsymbol{X}_i)\}\left[1 - g\{\boldsymbol{Y}_{Ci}^{\intercal}\boldsymbol{F}(\boldsymbol{X}_i)\}^{\beta}\right] k\left(g^{-1}\left[g\{\boldsymbol{Y}_{Ci}^{\intercal}\boldsymbol{F}^{(m-1)}(\boldsymbol{X}_i)\}g\{\boldsymbol{Y}_{Ci}^{\intercal}\boldsymbol{F}(\boldsymbol{X}_i)\}^{\beta}\right]\right),$$

with respect to $\boldsymbol{F} \in \mathcal{G}^*$ and $\beta \geq 0$, denoted by $\widehat{\boldsymbol{F}}$ and $\widehat{\beta}$. Then we update the classifier coordinate-wise as $F_j^{(m)} = g^{-1}\{g(F_j^{(m-1)})g(\widehat{F}_j)^{\widehat{\beta}})\}$ so that we have the following

$$g\{\boldsymbol{Y}_C^{\intercal}\boldsymbol{F}^{(m)}\} = g\{\boldsymbol{Y}_C^{\intercal}\boldsymbol{F}^{(m-1)}\}g\{\boldsymbol{Y}_C^{\intercal}\widehat{\boldsymbol{F}}\}^{\widehat{\beta}},$$

holding for all C. This will ensure that the property $\prod_{j=1}^{n} g(F_j^{(m)}) = 1$ continues to hold throughout the iterations. Thus at each iteration, we would be greedily maximizing:

$$\sum_{i=1}^{N}\left[g\{\boldsymbol{Y}_{Ci}^{\intercal}\boldsymbol{F}^{(m-1)}(\boldsymbol{X}_i)\} - g\{\boldsymbol{Y}_{Ci}^{\intercal}\boldsymbol{F}^{(m)}(\boldsymbol{X}_i)\}\right] k\left\{\boldsymbol{Y}_{Ci}^{\intercal}\boldsymbol{F}^{(m)}(\boldsymbol{X}_i)\right\},$$

which is exactly the empirical version of (2.3.3).

For illustration, consider $g(x) = e^x$ with ϕ being differentiable and hence we may let $k(x) = \dot{\phi}(x)e^{-x}$. In this special case the update of $\boldsymbol{F}^{(m)}$ simplifies to $\boldsymbol{F}^{(m)} = \boldsymbol{F}^{(m-1)} + \widehat{\beta}\widehat{\boldsymbol{F}}$. Therefore following the iteration described above we have:

$$\underset{\boldsymbol{F}\in\mathcal{G}^*,\,\beta\geq 0}{\operatorname{argmin}} \sum_{i=1}^{N} -\{e^{-\beta\boldsymbol{Y}_{Ci}^{\intercal}\boldsymbol{F}(\boldsymbol{X}_i)} - 1\}\dot{\phi}\{\boldsymbol{Y}_{Ci}^{\intercal}\boldsymbol{F}^{(m-1)}(\boldsymbol{X}_i) + \beta\boldsymbol{Y}_{Ci}^{\intercal}\boldsymbol{F}(\boldsymbol{X}_i)\}. \tag{2.3.4}$$

Note here, the apparent similarity between a coordinate descent (or gradient descent in a functional space) as proposed in [60] and [24], and the above iteration. Finally, we summarize the algorithm as follows.

Algorithm 1 Generic Boosting Algorithm

1. Set $\boldsymbol{F}^{(0)} = 0$;
2. For $m = 1, \ldots, M$:
   (a) Maximize

$$\sum_{i=1}^{N} g\{\boldsymbol{Y}_{Ci}^{\intercal}\boldsymbol{F}^{(m-1)}(\boldsymbol{X}_i)\}\left[1 - g\{\boldsymbol{Y}_{Ci}^{\intercal}\boldsymbol{F}(\boldsymbol{X}_i)\}^{\beta}\right] \times k\left(g^{-1}\left[g\{\boldsymbol{Y}_{Ci}^{\intercal}\boldsymbol{F}^{(m-1)}(\boldsymbol{X}_i)\}g\{\boldsymbol{Y}_{Ci}^{\intercal}\boldsymbol{F}(\boldsymbol{X}_i)\}^{\beta}\right]\right) \tag{2.3.5}$$

   with respect to $\boldsymbol{F} \in \mathcal{G}^*$, $\beta \geq 0$ to obtain $\widehat{\boldsymbol{F}}$ and $\widehat{\beta}$.
   (b) Update $\boldsymbol{F}^{(m)}$ coordinate-wise as $F_j^{(m)} = g^{-1}\{g(F_j^{(m-1)})g(\widehat{F}_j)^{\widehat{\beta}})\}$.
3. Output $\boldsymbol{F}^{(M)}$ and classify via $\operatorname{argmax}_j H_\phi(F_j^{(M)})$;

2.3.2 Numerical Convergence of the Algorithm when $g(\cdot) = \exp(\cdot)$

In this subsection we illustrate how algorithm 1 performs in finite samples, if we let it run until convergence (using potentially infinitely many iterations). We specifically study the properties of the iteration above in the case when $g(x) = \exp(x)$, or in other words we are concerned with the iteration given by (2.3.4). In addition we also want to explore the relationship between iteration (2.3.4)

and the following optimization problem:

$$\inf_{\boldsymbol{F}\in \operatorname{span}\mathcal{G}^*} \sum_{i=1}^{N} \phi(\boldsymbol{Y}_{C_i}^{\intercal}\boldsymbol{F}(\boldsymbol{X}_i)) \tag{2.3.6}$$

To this end we formulate the following:

Definition 2.3.3 (Looping closure)**.** *Let π be a permutation of the numbers $\{1,\dots,n\}$ into $\{\pi_1,\dots,\pi_n\}$. Consider the following "loop" functions, such that for all $i = 1,\dots,n$: $l^{(0)}(\pi_i) = \pi_i, l^{(1)}(\pi_i) = \pi_{i+1}$, $l^{(k)}(\cdot) = l^{(1)}(l^{(k-1)}(\cdot))$, where the indexing is $\mod n$, and $k = 1,\dots,n-1$[§]. We say that a classifier bag $\mathcal{G}$ is closed under "looping" if there exists a permutation π such that for all $G \in \mathcal{G}$ it follows that $l^{(k)} \circ G \in \mathcal{G}$ for all $k = 0,\dots,n-1$.*

In practice, closure under looping can easily be achieved if it is not already present, by adding the missing classifiers to the bag. Similar bag closures have been considered in the two class case in Mason et al. [60]. We start our discussion with the following proposition, providing a property of the algorithm, at its limiting points.

Proposition 2.3.4. *Suppose that the loss function ϕ is decreasing, continuously differentiable, bounded from below and satisfies (2.2.5) with $g = \exp$. Furthermore assume that, the classifier bag is closed under looping (see Definition 2.3.3). Then iterating (2.3.4), using possibly infinite amount of iterations until a limiting point $\boldsymbol{F}^{(\infty)}$ is reached, guarantees that the following condition holds:*

$$\sum_{i=1}^{N} \boldsymbol{Y}_{Ci}^{\intercal}\boldsymbol{F}(\boldsymbol{X}_i)\dot{\phi}(\boldsymbol{Y}_{Ci}^{\intercal}\boldsymbol{F}^{(\infty)}(\boldsymbol{X}_i)) = 0, \tag{2.3.7}$$

for all $\boldsymbol{F} \in \mathcal{G}^*$.

[§]Note that the loop functions depend on the permutation π, but we suppress this dependence for clarity of exposition.

Clearly, all examples of loss functions we considered satisfy the assumptions in Proposition 2.3.4. In the case when ϕ is convex, (2.3.7) shows that by iterating (2.3.4) we would arrive at an infimum of (2.3.6). This can be easily seen along the following lines. Assume that the $\widetilde{\boldsymbol{F}}$ is a point of infimum of (2.3.6) over the span of $\mathcal{G}^*$. By convexity of ϕ we have:

$$\sum_{i=1}^{N} \phi(\boldsymbol{Y}_{Ci}^{\intercal}\widetilde{\boldsymbol{F}}(\boldsymbol{X}_i)) - \sum_{i=1}^{N} \phi(\boldsymbol{Y}_{Ci}^{\intercal}\boldsymbol{F}^{(\infty)}(\boldsymbol{X}_i)) \geq \sum_{i=1}^{N} \boldsymbol{Y}_{Ci}^{\intercal}[\widetilde{\boldsymbol{F}}(\boldsymbol{X}_i) - \boldsymbol{F}^{(\infty)}(\boldsymbol{X}_i)]\dot{\phi}(\boldsymbol{Y}_{Ci}^{\intercal}\boldsymbol{F}^{(\infty)}(\boldsymbol{X}_i)) = 0.$$

The last equality follows from (2.3.7) and the fact that $\widetilde{\boldsymbol{F}} - \boldsymbol{F}^{(\infty)}$ is in the span of classifiers.

In the case when ϕ is not convex, condition (2.3.7) remains meaningful, though it doesn't guarantee convergence to the infimum. In order for us to relate condition (2.3.7) to equation (2.2.10) in the general (non-convex loss) case, and make it more intuitive, we consider a simple and illustrative example. We restrict our attention to the two class case ($n = 2$), but the example can easily be generalized.

First note that the classification rule (2.2.10) in the two class case with $g = \exp$ becomes $\operatorname{argmax}_{j\in\{1,2\}} \frac{\{\dot{\phi}(\boldsymbol{Y}_j^{\intercal}\boldsymbol{F}^{(\infty)}(x))\}^{-1}}{\{\dot{\phi}(\boldsymbol{Y}_1^{\intercal}\boldsymbol{F}^{(\infty)}(x))\}^{-1}+\{\dot{\phi}(\boldsymbol{Y}_2^{\intercal}\boldsymbol{F}^{(\infty)}(x))\}^{-1}}$. Consider a (disjoint) partition of the predictor support: $\mathcal{X}$ – $\mathcal{X}_1, \ldots, \mathcal{X}_B$. Construct classifiers based on that partition in the following manner:

$$G_b(x) = \begin{cases} 1, & \text{if } x \in \mathcal{X}_b \\ 2 & \text{otherwise} \end{cases},$$

and close them under looping. It is easily seen that, under this framework the vector $\boldsymbol{F}^{(\infty)}(x)$ is constant for $x \in \mathcal{X}_b$ for a fixed b. Denote this value with $\boldsymbol{F}_b^{(\infty)}$. Plugging in the b^{th} classifier in equation (2.3.7) we obtain: $N_b\dot{\phi}(\boldsymbol{Y}_1^{\intercal}\boldsymbol{F}_b^{(\infty)}) - (N - N_b)\dot{\phi}(\boldsymbol{Y}_2^{\intercal}\boldsymbol{F}_b^{(\infty)}) = 0$, where N_b is the number of observations correctly classified by the b^{th} classifier, or in other words: $\frac{\{\dot{\phi}(\boldsymbol{Y}_1^{\intercal}\boldsymbol{F}_b^{(\infty)})\}^{-1}}{\{\dot{\phi}(\boldsymbol{Y}_1^{\intercal}\boldsymbol{F}_b^{(\infty)})\}^{-1}+\{\dot{\phi}(\boldsymbol{Y}_2^{\intercal}\boldsymbol{F}_b^{(\infty)})\}^{-1}} =$

$\frac{N_b}{N}$,[¶] which evidently is an estimate of the probability that $\mathbb{P}(C = 1|\boldsymbol{X} \in \mathcal{X}_b)$, which in turn is a proxy to the Bayes classifier. Moreover note that this completely agrees with the classification rule yielded by equation (2.2.10).

Convergence Analysis

In the convex loss function case, property (2.3.7) will be matched by a gradient descent methods in the function space (such as AnyBoost[60] e.g.). This motivates us to consider the question of the convergence rate of the newly suggested algorithm — is it slower, faster or the same as a gradient descent in the convex loss function case? At first glance the rate might appear to be slower as we are not using the "fastest" decrease at each iteration using simply the exp function. In the end of this subsection we establish a geometric rate of convergence under certain assumptions, which matches the convergence rate for gradient descent under similar assumptions.

As we argued in the previous subsection, in the case of a convex loss ϕ, (2.3.7) guarantees that iteration (2.3.4) converges to the infimum of problem (2.3.6). Let $\boldsymbol{Y}_{C_i}^{\mathsf{T}} \boldsymbol{F}^{(\infty)}(\boldsymbol{X}_i)$ be the limiting (allowed to be $\pm\infty$) values achieving the infimum above. Before we formalize the convergence rate result, we will characterize the behavior of $\boldsymbol{Y}_{C_i}^{\mathsf{T}} \boldsymbol{F}^{(\infty)}(\boldsymbol{X}_i)$.

This question is of interest in its own right, as this characterization remains valid regardless of what boosting algorithm one decides to use to obtain the minimum/infimum. For what follows we consider a loss function ϕ, which satisfies a mildly strengthened condition (2.2.11). Namely, let ϕ be decreasing and for any $\alpha, c > 0$ it satisfies the following condition:

$$\phi(x) + c\phi(-\alpha x) \uparrow +\infty \text{ as } x \uparrow +\infty \tag{2.3.8}$$

[¶] Here we assume that $\dot{\phi}(\boldsymbol{Y}_j^{\mathsf{T}} \boldsymbol{F}_b^{(\infty)}) \neq 0, j \in \{1, 2\}$, which can be ensured if ϕ is unbounded from above

It is worth noting that if condition (2.2.11) is satisfied for all n (recall that $g = \exp$ here) this would imply (2.3.8). Denote with B the total number of weak learners in the bag. Let $\boldsymbol{D} = \{\boldsymbol{Y}_{Ci}^{\intercal}\boldsymbol{F}_j(\boldsymbol{X}_i)\}_{j,i}$ be the $B \times N$ matrix, each entry of which is either $\mathcal{C}_+$ or $\mathcal{C}_-$. Again, we assume that the bag is closed under looping. Let $\boldsymbol{v} \in \mathbb{R}^N$ be a vector. Consider the equation $\boldsymbol{D}^{\intercal}\boldsymbol{\alpha} = \boldsymbol{v}$ for some vector $\boldsymbol{\alpha} \in \mathbb{R}^B_{0,+}$ with non-negative coordinates. Note that because of the looping closure[‖] the linear equation above has solution iff the equation $\boldsymbol{D}^{\intercal}\boldsymbol{\alpha} = \boldsymbol{v}$ has a solution with $\boldsymbol{\alpha} \in \mathbb{R}^B$, since without loss of generality we can add a large positive constant to the coordinates of $\boldsymbol{\alpha}$. It follows that the equation $\boldsymbol{D}^{\intercal}\boldsymbol{\alpha} = \boldsymbol{v}$ with $\boldsymbol{\alpha} \in \mathbb{R}^B_{0,+}$ has a solution iff $\boldsymbol{v} \in \mathrm{row}(\boldsymbol{D})$.

To see the connection between the linear equation above and optimization problem (2.3.6) consider the following simple example. The function $\sum_{i=1}^N \phi(\sum_{j=1}^B \alpha_j \boldsymbol{Y}_{Ci}^{\intercal}\boldsymbol{F}_j(\boldsymbol{X}_i))$ cannot have a minimum, if there exists a vector $\boldsymbol{v} \in \mathbb{R}^N_+$ with strictly positive coordinates, such that the equation $\boldsymbol{D}^{\intercal}\boldsymbol{\alpha} = \boldsymbol{v}$ has a solution — $\widehat{\boldsymbol{\alpha}} \in \mathbb{R}^B_{0,+}$. To see this, suppose the contrary, take an arbitrary constant $R > 0$ and note that:

$$\sum_{i=1}^N \phi(\sum_{j=1}^B R\widehat{\alpha}_j \boldsymbol{Y}_{Ci}^{\intercal}\boldsymbol{F}_j(\boldsymbol{X}_i)) = \sum_{i=1}^N \phi(Rv_i).$$

Take the limit $R \to \infty$, and it is clear that the infimum $N\phi(+\infty)$ is achieved. It follows that if we want to have a solution smaller than $N\phi(+\infty)$ — $\boldsymbol{D}$ cannot have rank N. Denote the rank of $\boldsymbol{D}$ with r.

More generally, our next result provides a characterization of how many (and which) of the values $\boldsymbol{Y}_{C_i}^{\intercal}\boldsymbol{F}^{(\infty)}(\boldsymbol{X}_i)$ are set to $+\infty$ at the infimum of (2.3.6). Consider the perp space of the row space of the matrix $\boldsymbol{D}$ — $\boldsymbol{E} := \mathrm{row}(\boldsymbol{D})^{\perp}$. Out of all possible bases of $\boldsymbol{E}$ including the 0 vector, select the one $\boldsymbol{e}_1, \ldots, \boldsymbol{e}_s$ ($s = \min(N, B) - r + 1$) for which the vector $\boldsymbol{e}_1$ has the most strictly positive entries at I coordinates and zeros at the rest[**]. We have the following:

[‖] Looping closure (Definition 2.3.3) gives us the the column sums of $\boldsymbol{D}$ are 0.

[**] We allow $I = 0$, in which case $\boldsymbol{e}_1$ would simply represent the 0 vector.

Proposition 2.3.5. *Let ϕ be a decreasing loss function satisfying (2.3.8). Set $M := N - I$, where $I \in \{0, \dots, N\}$. We have that:*

$$(N - M - 1)\phi(0) + (M + 1)\phi(+\infty) < \inf_{\boldsymbol{F} \in \operatorname{span} \mathcal{G}^*} \sum_{i=1}^{N} \phi(\boldsymbol{Y}_{C_i}^\intercal \boldsymbol{F}(\boldsymbol{X}_i)) \leq (N - M)\phi(0) + M\phi(+\infty).$$

Moreover, exactly M of the values $\boldsymbol{Y}_{C_i}^\intercal \boldsymbol{F}^{(\infty)}(\boldsymbol{X}_i)$ (i will be corresponding to the 0 coordinates of $\boldsymbol{e}_1$) will be set to $+\infty$ at the infimum.

Proposition 2.3.5 characterizes the cases when one should expect problem (2.3.6) to have a minimum. In fact, in the cases where $I > 0$, we can simply delete the observations corresponding to the rows of $\boldsymbol{e}_1$ that are 0, and solve the optimization only on the set of observations left, as it can be seen from the proof.

We next formulate the speed of the convergence of the algorithm we suggested, in the case when the function ϕ is convex. For simplicity we assume that the matrix of classifier entries — $\boldsymbol{D}$, is such that there is a strictly positive vector in the perp of the row space of $\boldsymbol{D}$. If that is not the case as argued we can delete observations that will be set to $+\infty$ at the maximum, and work with the rest. Denote with $\mathcal{S} = \{\boldsymbol{v} : \boldsymbol{D}^\intercal \boldsymbol{\alpha} = \boldsymbol{v} \text{ with } \boldsymbol{\alpha} \geq 0, \sum_{i=1}^{N} \phi(v_i) \leq N\phi(0)\}$. Proposition 2.3.5 then implies that, the set $\mathcal{S}$ is bounded coordinate-wise. Next we formulate the result:

Theorem 2.3.6. *Let the convex, decreasing loss function ϕ be strongly convex with Lipchitz and bounded derivative on any compact subset of $\mathbb{R}$, and satisfies (2.3.8) and (2.2.5) with $g = \exp$. Furthermore, assume that there is a stricly positive vector in $\operatorname{row}(\boldsymbol{D})^\perp$, and define the set $\mathcal{S}$ as above. Let $\boldsymbol{F}^* \in \operatorname{span} \mathcal{G}^*$ achieves the minimum in problem (2.3.6). Denote with $\varepsilon_m = \sum_{i=1}^{N} \phi(\boldsymbol{Y}_{Ci}^\intercal \boldsymbol{F}^*(\boldsymbol{X}_i)) - \sum_{i=1}^{N} \phi(\boldsymbol{Y}_{Ci}^\intercal \boldsymbol{F}^{(m)}(\boldsymbol{X}_i))$, where $\boldsymbol{F}^{(m)}$ is produced iteratively using (2.3.4). Then there exists a con-*

stant $K < 1$ *depending on the matrix* $\boldsymbol{D}$*, the sample size* N *and the set* $\mathcal{S}$*, such that:*

$$\varepsilon_{m+1} \leq \varepsilon_m K.$$

As we can see from Theorem 2.3.6 if we use this algorithm in the convex loss function case, we wouldn't lose convergence speed to gradient descent (see[67] Theorem 2.1.14), but in the non-convex function case which still obeys (2.2.5) this algorithm will be converging to a local minimum. In the latter case we will still be capable of recovering the Bayes classifier, as indicated by equation (2.3.7).

2.3.3 Aggregating Boosted Classifiers Via Cross-validation

The performance of the boosting algorithm is likely to be dependent on the choice of ϕ for a given dataset. An optimal ϕ can be selected via procedures such as the CV. On the other hand, optimally combining information from multiple boosting algorithms trained with different ϕ to further improve the robustness of our predictions in terms of outliers would be valuable. We propose a simple approach to address this by optimally combining predicted probabilities recovered from multiple boosting algorithms as illustrated in (2.2.10). Similar CV based aggregation approach has been previously proposed to select or linearly combine multiple learners to optimize an L_2 loss[74,81].

Let $\omega_j^{[\ell]}(\boldsymbol{X})$ denote the estimate of $\omega_j(\boldsymbol{X}) = \text{logit}P(C = j \mid \boldsymbol{X})$ based on the boosting algorithm with the ℓ^{th} loss function, for $\ell = 1, ..., L$, where L is the total number of losses under consideration. Then an improved estimate of $\omega_j(\boldsymbol{X})$ can be obtained by fitting a multinomial regression with C_i being the outcome and $\{\omega_j^{[\ell]}(\boldsymbol{X}_i), j = 1, ..., n-1, \ell = 1, ..., L\}$ being the predictors. To overcome over-fitting and potentially high collinearity between $\{\widehat{\omega}_j^{[\ell]}(\boldsymbol{X})$ and $\widehat{\omega}^{[\ell']}(\boldsymbol{X})$ when $\ell \neq \ell'$, we employ the CV with a simple ridge regularization in the multinomial regression fitting. Specifically, we partition the data into $\mathcal{K}$ parts, $\{\mathcal{D}^{(\kappa)}, \kappa = 1, ..., \mathcal{K}\}$. For $\kappa = 1, ..., \mathcal{K}$, we use data not in $\mathcal{D}^{(\kappa)}$ to train the L algorithms and obtain the corresponding $\{\widehat{\omega}_j^{[\ell]}(\cdot)\}$, denoted

by $\{\widehat{\omega}^{[\ell]}_{j(-\kappa)}(\cdot)\}$. Then combining initial fittings from all $\mathcal{K}$ partitions, we construct synthetic data with C_i being the outcome and $\boldsymbol{\varpi}_i = \{1, \widehat{\omega}^{[\ell]}_{j(-\kappa_i)}(\boldsymbol{X}_i), j = 1, ..., n-1, \ell = 1, ..., L\}$ being the covariate vector, where κ_i is the index such that the ith observation belongs to $\mathcal{D}^{(\kappa_i)}$. We fit a multinomial ridge regression $P(C_i = \jmath \mid \boldsymbol{\varpi}_i) = g_{\text{logit}}(\boldsymbol{\gamma}_\jmath^\intercal \boldsymbol{\varpi}_i)$ with the synthetic data and obtain coefficients $\widehat{\boldsymbol{\gamma}}_j$, where g_{logit} is the anti-logit function. The final classification combining information from all L algorithms is then based on $\text{argmax}_\jmath \{g(\widehat{\boldsymbol{\gamma}}_\jmath^\intercal \boldsymbol{\varpi})\}$.

2.4 Numerical Studies and Data Example

In this section we validate empirically the performance of the generic boosting algorithm developed in the previous section, comparing it to popular classification algorithms such as SVM and SAMME on synthetic data. We further apply the algorithm to a electronic medical record study on diabetic neuropathy conducted at the Partners Healthcare.

2.4.1 Simulation Studies

We conducted simulation studies to evaluate the performance of our proposed procedures compared to existing methods and examine how the choice of ϕ may impact the classification accuracy. For each dataset generated from each of the configuration described below, we evaluated our proposed boosting algorithm based on (i) $\phi(x) = \log(1 + e^{-x})$ (Logistic) with $g(x) = e^{cx}$ and $k(x) = -\{ce^{cx}(1 + e^x)\}^{-1}$, for $c = 0.1$; (ii) $\phi(x) = \log(\log(e^{-x} + e))$ (LogLog) with $g(x) = e^x$ and $k(x) = \{e^x(e^{x+1} + 1)\log(e^{-x} + e)\}^{-1}$. We also compare each of these algorithms to the CV aggregated algorithm (CV) as well as to the commonly used LASSO and SVM procedures. The SVM was trained with RBF kernel where the tuning parameter for the kernel function was chosen via the `sigest` function of `ksvm` library. The `sigest` procedure outputs three quantiles – 0.1, 0.5, 0.9 of the distribution of $||X - X'||^2$ where X and X' are two predictors sampled

from the matrix $\boldsymbol{X}$, and we take the mean of these quantiles as the tunning parameter in the RBF kernel for robustness. The fitting was performed with the `kernlab` `R` package implementation – `ksvm` which uses the "one-against-one"-approach to deal with multi-class problems see[34] for example. The LASSO procedure with $\boldsymbol{X}$ being the predictors was based on an ℓ_1 penalized continuation ratio logistic regression[4], where the tuning was selected based on 5-fold CV.

Across all configurations, we generate $\boldsymbol{X} = (X_1, ..., X_{50})^\intercal$, so that each X_i is marginally distributed as $U(-1, 1)$, and overall they have exchangeable correlation, with the off diagonal of the correlation matrix being .4. To achieve this, we first generate normal variables $Z \sim N(0, \Theta)$ and we invert them by applying $\boldsymbol{X} = F(Z)$ coordinate-wize, where F is the cdf of the standard normal distribution. We use a sample size of $N_t = 200$ for training and $N_v = 3000$ for independent validation. A large validation size is chosen so that the variation observed in the classification performance in the validation set is reflecting the variability of different algorithms obtained with the training set. All boosting algorithms were performed based on 50 iterations. Across all settings, we let $n = 3$ for generating the outcome and summarize results based on 50 replications.

For a given $\boldsymbol{X}$, we generate C from multinomial with success probabilities for $C = 1, C = 2$ being $p_1 = .7 - .6 \min(R^4, 1)$ and $p_2 = .1 + .6 \min(R^4, 1)$, respectively, where we consider 5 different scenarios for choosing R:

$$R = \begin{cases} (1 + \text{sign}(X_1 - 1/3))/2, & \text{Setting (I)} \\ \text{sign}(\text{acos}(X_2)/\pi - 1/4))/3 + (1 + \text{sign}(X_1 - 1/3))/3, & \text{Setting (II)} \\ \frac{1}{2} + \cos(X_2^2)/4 + \text{sign}(X_1 - \frac{1}{3})/4, & \text{Setting (III)} \\ \frac{3}{4} + \cos(X_2^2)/4 + \text{sign}(X_1 - \frac{1}{3})/4 + 1/8 \sum_{d=3}^{10} \text{sign}(X_d - 1/4), & \text{Setting (IV)} \\ \frac{1}{2} + \cos(X_2^2)/4 + \text{sign}(X_1 - \frac{1}{3})/4 + 1/8 \sum_{d=3}^{10} X_d, & \text{Setting (V)} \end{cases}$$

The signals are very sparse and non-linear in settings I-III. Settings in IV and V are less sparse with setting IV being mostly non-linear and setting V being a mixture of linear and non-linear signals. Across all these settings, $\boldsymbol{X}$ is unrelated to the probability of being in class 3.

For all the boosting algorithms, we create a bag of weak learners based on "*decision stumps*", where for each predictor variable X_d, we choose a sequence of threshold values $-1 < x_{d1} < \cdots < x_{dM} < 1$ and for each pair $\{x_{dm}, x_{dm'}\}$, we create a classification of C based on the three regions defined by $\{x_{dm}, x_{dm'}\}$. To improve the classification and computation efficiency, the bag for constructing the boosting algorithms consists of all decision stumps that yield at least 45% correct classifications on the training set.

To quantify the performance of each of the algorithm, we use the misclassification error rate relative to that of the oracle Bayes rule. The average relative error along with the standard deviation of the error rates across 50 replications are reported in Table 2.1. Across all configurations, the Logistic and LogLog losses from our proposed algorithms perform better than the SAMME algorithm proposed in[100]. Interestingly, the LogLog loss performs the best among all three losses with lower misclassification rates and lower variability. This could in part due to the fact that the non-convexity of the LogLog loss is less sensitive to outliers[59]. In addition, our proposed CV aggregation procedure seems to perform well in combining information from all 3 losses, producing classifications that are almost always at least as accurate as those from the best of the 3 boosting algorithms. Comparing to the LASSO and SVM, our boosting algorithms based on the LogLog loss or CV aggregation always outperform these commonly used methods. This could in part be due to the fact that the signals are sparse and non-linear, under which case neither LASSO or SVM are expected to work well.

2.4.2 Data Example

To illustrate our proposed generic boosting algorithm and demonstrate the advantage of having multiple losses, we apply our procedures to an electronic medical record (EMR) study, conducted at

Table 2.1: The average (standard deviations) relative misclassification rate for the 5 settings. The misclassification rates are 30%, 31%, 35%, 34% and 35% for settings I, II, III, IV, and V, respectively.

	Setting I	Setting II	Setting III	Setting IV	Setting V
SVM	1.43 (0.09)	1.48 (0.07)	1.29 (0.05)	1.30 (0.06)	1.29 (0.07)
LASSO	1.17 (0.07)	1.30 (0.10)	1.14 (0.08)	1.15 (0.06)	1.18 (0.10)
SAMME	1.24 (0.11)	1.37 (0.08)	1.18 (0.06)	1.21 (0.07)	1.21 (0.08)
Logistic	1.19 (0.08)	1.29 (0.06)	1.15 (0.05)	1.18 (0.07)	1.18 (0.08)
LogLog	1.10 (0.06)	1.20 (0.05)	1.09 (0.05)	1.11 (0.06)	1.12 (0.08)
CV	1.06 (0.06)	1.19 (0.07)	1.09 (0.06)	1.11 (0.06)	1.13 (0.12)

the Partners Healthcare, aiming to identify patients with different subtypes of diabetic neuropathy. Diabetic neuropathy (DN), a serious complication of diabetes, is the most common neuropathy in industrialized countries[71]. It is estimated that 20 million people worldwide are affected by symptomatic diabetic neuropathy. Growing rates of obesity and the associated increase in the prevalence of type 2 diabetes could cause these figures to double by the year 2030. The prevalence of DN also increases with time and poor glycemic control[58]. Although many types of neuropathy can be associated with diabetes, the most common type is diabetic polyneuropathy and pain can develop as a symptom of diabetic polyneuropathy[77,26]. Pain in the feet and legs was reported to occur in 11.6% of insulin dependent diabetics and 32.2.1% of noninsulin dependent diabetics[102]. Unfortunately, risk factors for developing painful diabetic neuropathy (PDN) are generally poorly understood. PDN has been reported as more prevalent in patients with type 2 diabetes and women[1]. Prior studies have also reported an association between family history and PDN, suggesting a potential genetic predisposition to PDN[26].To enable a genetic study of PDN and non-painful DN (nPDN), an EMR study was performed to identify patients with these two subtypes of DN by investigators from the informatics for integrating biology to the bedside (i2b2), a National Center for Biomedical Computing based at Partners HealthCare[65,66].

To identify such patients, we created a datamart compromising 20,000 patients in the Partners

Healthcare with relevant ICD9 (International Classification of Diseases, version 9) codes. Two sources of information were utilized to classify patients' DN status and subtypes: (i) structured clinical data searchable in the EMR such as ICD9 codes; and (ii) variable identified using natural language processing (NLP) to identify medical concepts in narrative clinical notes. Algorithm development and validation was performed in a training set of 611 patients sampled from the datamart. To obtain the gold standard disease status for these patients, several neurologists performed chart reviews and classified them into no DN, PDN and nPDN. To train the classification algorithms, we included a total of 85 predictors most of which are NLP variables, counting mentions of medical concepts such as *"pain", "hypersensitivity"*, and *"diabetic neuropathy"*.

We trained boosting classification algorithms to classify these 3 disease classes. We used simple decision trees as weak learners. They only have two nodes with the first node deciding between class C_1 vs C_2 and C_3 and the other node deciding between C_2 vs C_3, where $\{C_1, C_2, C_3\}$ is a permutation of {noDN,PDN,nPDN}. In order to illustrate the algorithms we sample 350 observations and use them as a training set and the rest 361 patients we set off as a test set.

We report the percentage mis-classifications:

Table 2.2: Percent mis-classifications

	% incorrect
SVM	32%
LASSO	30%
SAMME	27%
Logistic	26%
LogLog	26%
CV	27%

The boosting results show a modest improvement, as compared to standard methods. We can also see that the generic boosting algorithm performs slightly better than SAMME in this situation

with both the logistic and the loglog losses. It warrants further research whether picking richer tree structures would yield an even better performance on this dataset.

2.5 Discussion

For multi-category classification problems, we described in this paper a class of loss functions that attain FC properties and provided theoretical justifications for how such loss functions can ultimately lead to optimal Bayes classifier. We extended the results to accommodate differential costs in misclassifying different classes. To approximate the minimizer of the empirical losses, we demonstrated that a natural iterative procedure can be used to derive generic boosting algorithms for any of the proposed losses. To further improve the robustness of the proposed boosting algorithms, we proposed a CV based aggregation procedure to combine information from boosting classifiers from multiple losses. Simulation results suggest that non-convex losses could potentially lead to algorithms with better performance and our CV aggregated algorithm almost always achieve the lowest error rate when compared to other boosting algorithms.

Our proposed algorithm not only depends on the choice of ϕ but also the associated $g(\cdot)$ and $k(\cdot)$ functions as indicated in (2.2.5). We can think of g as a positive deformation of the real line and even with the same ϕ, changing g could also change the classifiers. Most existing boosting algorithms correspond to $g(x) = e^x$, in which case the constraint $\prod_{j=1}^n g(F_j) = 1$ simplifies to the commonly seen condition $\sum_j F_j = 0$. Moreover if ϕ is smooth and convex, one may let $k(x) = \dot{\phi}(x)/e^x$. Thus, under convexity, $H_\phi(x) = \dot{\phi}(x) = d\phi(x)/dx$ is an increasing function and ϕ is Fisher consistent in the traditional sense. We also saw, that even when ϕ is not convex, our suggested losses are Fisher consistent in the standard sense. Moreover, we argued that loss functions satisfying (2.2.5), can be used to recover the exact conditional probabilities. It would be interesting to develop adaptive boosting procedure where we use different g functions in the process of boost-

ing adaptively. For example, in the suggested logistic loss boosting with $g(x) = e^{cx}$, we can adaptively select the parameter c, for better convergence results of the algorithm which will potentially result in a better classification results. We were provided a property of the limiting point of the algorithm in the case where $g = \exp$. Furthermore, we characterized when the problem has a minimum in the finite sample case under certain assumptions on ϕ. The resemblance of the proposed generic boosting algorithm with coordinate descent, helped us to establish geometric rate of convergence in the convex loss function case. The consistency of the algorithm under conditions such as finite VC dimension of the classifier bag, warrants future research.

He uses statistics as a drunken man uses lamp-posts—for support rather than illumination.

Andrew Lang

3

Support Recovery for Sliced Inverse Regression in High Dimensions

3.1 Introduction

In this chapter we study the Sliced Inverse Regression (SIR) procedure in a high-dimensional setting. The SIR was suggested in the seminal paper[46]. SIR is the supervised counterpart of principal

component analysis (PCA)[57,5]. SIR is a dimension reduction tool, projecting the data onto a lower dimensional space, but in contrast to PCA, SIR leverages information for an outcome of interest Y. The original SIR procedure was designed to handle models of the sort:

$$Y = f(\beta_1^\intercal X, \beta_2^\intercal X, \ldots, \beta_r^\intercal X, \varepsilon),$$

where β is a p-dimensional vector, ε is random noise independent of X, f is an unknown function, and r is the number of linear components participating in the model. If r turns out to be small compared to p we would gain insights for the data if we are able to estimate the vectors $\beta_i, 1 \leq i \leq r$, by a projection of the predictor on the sufficient dimension reduction (SDR) space $=$ $\text{span}\{\beta_1, \ldots, \beta_r\}$. SIR operates by slicing the outcome Y into cuts, averaging the predictors, and performing a singular value decomposition on the weighted conditional covariance matrix. Under certain assumptions SIR provably recovers the SDR space, in the low dimensional regime when $p < n$, e.g. see [46,33] . The most notable of the assumptions required for SIR to work is the assumption of linearity in expectation, i.e. $\mathbb{E}[b^\intercal X | \beta_1^\intercal X, \ldots, \beta_r^\intercal X] = \sum_{i=1}^r c_i \beta_i^\intercal X_i$ for any b, or in other words the conditional expectation for any direction b is linear in terms of the projections on SDR directions. This property is satisfied by all elliptical distribution families.

Recently many papers studying sparse PCA procedures have emerged, starting with the seminal papers by Johnstone and Lu [37],[38], where the authors showed that PCA can be inconsistent in the regime $p/n \to c > 0$. This analysis was further strengthened in [70], where a stronger inconsistency result appeared. This justified the need to consider scenarios where the principal eigenvectors are sparse. The algorithm Diagonal Thresholding (DT) was suggested by [37], to deal with the spiked-covariance model. It was later analyzed in [3] to show that support recovery is achieved by DT in the sparse PCA spiked covariance model case when $n \gtrsim s^2 \log(p)$. In [3] the the authors further showed an information theoretic obstruction showing that no algorithm can recover the support of the

principle eigenvector if $n \lesssim s\log(p)$. An algorithm that succeeds in support recovery with high probability as long as $n \gtrsim s\log(p)$, but is computationally prohibitive, is exhaustively scanning through all $\binom{p}{s}$ subsets of the coordinates of the principal eigenvector. For that purpose, in [3] the authors studied a semidefinite programming (SDP) estimator originally suggested by d'Aspremont et al. [17] – and showed that under the assumptions $n \gtrsim s\log(p)$ and if the SDP has a rank 1 solution this solution can recover the signed support with high probability. Surprisingly, however [41], showed that the rank 1 condition, does not hold if $s^2\log(p) \gtrsim n \gtrsim s\log(p)$.

In the SIR literature, when p is fixed, the first asymptotic results appeared in the important paper by Hsing and Carroll [33]. Later on, p was allowed to diverge slowly with n, e.g. when $p = o(n^{1/2})$ asymptotic results were established in Zhu et al. [101]. In the super high-dimensional setting where $p \gg n$, several algorithms, hinging on regularization such as LASSO [79] and Dantzig Selector [14] were proposed by Li and Nachtsheim [48], Yu et al. [93], but these algorithms are not concerned with support recovery. Moreover, the algorithm suggested in [48] did not come with theoretical guarantees, and in Yu et al. [93] the authors did not allow s to increase with p and n. A generic variable selection procedure, was suggested in [99], with guarantees of support recovery, in a more general setting than our presentation in this chapter, but with a much more restrictive relationship $p = o(n^{1/2})$ than the one we consider.

In this chapter we study the DT and SDP algorithms applied to SIR and show that in fact both algorithms can achieve support recovery as long as $n \gtrsim s\log(p)$, in contrast to the PCA case, and furthermore we show an information theoretic obstruction as in Amini and Wainwright [3] which shows that in fact support recovery with high probability is impossible in the case when $n \lesssim s\log(p)$. This implies that in the SIR setting, the computational and statistical tradeoffs phenomenon does not appear in contrast to the PCA case. To the best of our knowledge, we provide the first result in the SIR literature allowing the sparsity s to diverge with p and n. We also provide numerical studies which confirming our findings.

3.1.1 Setup and Notation

In this chapter we are only concerned with a SIR setup with one dimensional SDR, or in other words our model takes the form:

$$Y = f(\beta^\intercal X, \varepsilon), \tag{3.1.1}$$

where the error distribution ε is independent of X. This setup is related to the single index model. Sparse single index models have been considered in Alquier and Biau [2]. Our framework is different from the one considered by Alquier and Biau [2], in many ways, most notably — our model is more generic, and we are interested in recovering the support of β, whereas Alquier and Biau [2] are concerned with estimation, and therefore the methods considered by us are completely different than the methods in [2]. Similarly to other papers concerned with support recovery in the PCA setting, e.g. [3,19], we assume a stylized setting with β being a p-dimensional sparse unit vector* with $\beta_i = \pm\frac{1}{\sqrt{s}}$ for $i \in \text{supp}(\beta) = \{i : \beta_i \neq 0\}$, for some $s \in \mathbb{N}$. We will use $S_\beta := \text{supp}(\beta)$ and $S_\beta^c := \text{supp}(\beta)^c = \{i : \beta_i = 0\}$, as a shorthand notations. Since β is a unit vector, we clearly have $|S_\beta| = s$ and $\beta_i = 0$ for $i \in S_\beta^c$. We further assume that $X \sim N(0, \mathbb{I}_{p\times p})$. While the latter is a rather simplifying assumption, we believe it is an important first step for studying the support recovery in SIR without the complications of how a covariance structure would modify the of sparsity in the β vector. Note that such X satisfies the linearity condition trivially. We observe n samples from this model, with potentially $n < p$ and even $n \ll p$. We are interested in studying approaches inspired by the SIR procedure, for support recovery and estimation purposes.

The procedures we study are concerned with the scenario where we slice the support of Y in H equally sized slices, with m observations in each slice. In other words, if $\{Y_{(1)}, \ldots, Y_{(n)}\}$ are the order statistics of the Y sample, the h^{th} slice consists of the observations with Y values in the set —

*Without loss of generality, we consider β being a unit vector for identifiability.

$\{Y_{((h-1)m+1)}, \ldots, Y_{((h-1)m)}\}$. Equivalently, the h^{th} slice consists of points, whose Y values are located in the following interval $(Y_{((h-1)m)}, Y_{(hm)}]$. Denote for the ease of notation the points in the h^{th} slice as $(Y_{h,i}, X_{h,i})$ with $i = 1, \ldots, m$, where $Y_{h,i} = Y_{((h-1)m+i)}$ (we use the notation for Y interchangeably). Potentially removing several observations at random, we can always assume for simplicity that $n = mH$. We will denote with superscript coordinates of the predictor vectors, with subscripts reserved to indicate, slice and observation indications.

The classical SIR procedure relies on constructing the conditional covariance matrix for the within sliced means estimator — V, where the j, k^{th} element of V is given by V^{jk}:

$$V^{jk} = \frac{1}{H} \sum_{h=1}^{H} \left(\frac{1}{m} \sum_{i=1}^{m} X_{h,i}^{j} \right) \left(\frac{1}{m} \sum_{i=1}^{m} X_{h,i}^{k} \right).$$

Note that since our data is assumed to be centered at 0, we do not need to further center at this step. This differs slightly with the originally proposed SIR estimate, which centers the data, but we show that we can handle that case as well (see e.g. Corollary 3.2.4).

We proceed with defining several helpful notations, which will make the presentation easier later on. Let $\overline{X}_{h,S}^{j} = \frac{1}{|S|} \sum_{i \in S} X_{h,i}^{j}$, where $S \subset \{1, 2, \ldots, m\}, j = 1, \ldots, p$. If $S = \{1, 2, \ldots, m\}$ we omit it from the notation, i.e. $\overline{X}_{h}^{j} = \frac{1}{m} \sum_{i=1}^{m} X_{h,i}^{j}$.

In terms of this notation we can rewrite the variance estimator as:

$$V^{jk} = \frac{1}{H} \sum_{h=1}^{H} \overline{X}_{h}^{j} \overline{X}_{h}^{k}. \tag{3.1.2}$$

Let $S_h = (Y_{((h-1)m)}, Y_{(hm)}], 1 \leq h < H$, and $S_H = (Y_{((H-1)m)}, +\infty)$, denote the random intervals whose end points are the $(h-1)m$ and hm order statistics of the Y sample correspondingly (with $Y_{(0)} = -\infty$). Denote with $\mu_h^j = \mathbb{E}[X^j | Y \in S_h]$. Furthermore let $m_j(Y) = \mathbb{E}[X^j | Y]$ denote the j^{th} coordinate of the so-called centered inverse regression curve. To this end note that

conditionally on the values $Y_{((h-1)m)}$ and $Y_{(hm)}$ the quantities S_h and μ_h^j become constants.

If $M \in \mathbb{R}^{d\times d}$ is a matrix, by double indexing with two subsets $S_1 \subset \mathbb{R}^d$ and $S_2 \subset \mathbb{R}^d$ — M_{S_1,S_2}, we mean taking the sub matrix corresponding to entries M_{ij} with $i \in S_1, j \in S_2$. Furthermore, we will use several different norms of vectors and matrices, which are briefly defined below. For a vector v, let $||v||_p$ denote the usual ℓ_p norm for $1 \leq p \leq \infty$ (using the usual extension for $p = \infty$), and by $||v||_0$ we denote $|\operatorname{supp}(v)|$.

Furthermore, for a $d \times d$ matrix $M_{d\times d}$, let $||M||_{\max} = \max_{jk} |M_{jk}|$ denote the entry-wise sup norm. Moreover, let $||M||_{p,q} = \sup_{||v||_p=1} ||Mv||_q$ denote the ℓ_p and ℓ_q induced norm on M. In particular in the special cases when $p = q = 2$ and $p = q = \infty$, we have:

$$||M||_{2,2} = \max_{i=1,\ldots,d} \{\sigma_i(M)\},$$

where $\sigma_i(M)$ represents the i^{th} singular of M, and:

$$||M||_{\infty,\infty} = \max_{i\in\{1,\ldots,d\}} \sum_{j=1}^{d} |M_{ij}|.$$

For a real valued random variable X, define the following Orlicz norms:

$$||X||_{\psi_2} = \sup_{d\geq 1} d^{-1/2}[\mathbb{E}|X|^d]^{1/d}, \tag{3.1.3}$$

$$||X||_{\psi_1} = \sup_{d\geq 1} d^{-1}[\mathbb{E}|X|^d]^{1/d}. \tag{3.1.4}$$

Finally, let $F_n(x) = \frac{1}{n}\sum_{i=1}^n \mathbb{I}(Y_i \leq x)$ denote the empirical distribution of the Y sample. We also use the standard notations Φ and ϕ to refer to the cdf and pdf of a standard normal random variable.

Chapter 3 is organized as follows: in Section 3.2 we present our main results, in Section 3.3 we

present numerical confirmations of the predictions of our main results, in Section 3.2.1 we analyze some of the conditions required for the main results, and in the next Sections — Section 3.4, 3.5 we show our main results. In Section 3.6 we study a support recovery from a slightly different angle, allowing us to generalize the assumption $\Sigma = \mathbb{I}$, and show the consistency of the linear regression LASSO algorithm, provided that certain restrictions on the covariance Σ are met, most notably the irrepresentable condition. Finally Section 3.7 is left for a brief discussion. Some of the technical arguments are deferred to Appendix B.

3.2 Main Results

In this section we consider two procedures for support recovery, and show the asymptotic consistency of the procedures, under the assumption that $\frac{n}{s\log(p-s)} > \Omega$ for a large enough Ω. Furthermore, we derive a lower bound on the sample size as a function of the sparsity s and the dimension p under which scenario, support recovery with high-probability is impossible. Before we go to the procedures we formulate some technical assumptions which we will need. We comment on the achievability of these assumptions in Section 3.2.1. Though we could have directly imposed the sufficient conditions provided in Section 3.2.1 instead of looking into the somewhat convoluted assumptions below, we believe that making these assumptions makes the intuition more explicit. Throughout the rest of the chapter we assume that Y is a continuously distributed random variable (exception being Section 3.6).

We proceed to define an assumption on the inverse regression curve:

Definition 3.2.1. *We call the pair (f, ε) sliced stable iff there exist constants $l < 1, K > 1, M > 0$, such that for any $H \in \mathbb{N}, H > M$, and all partitions of $\mathbb{R} = \{a_1 = -\infty, \dots, a_{H+1} = +\infty\}$ with $\frac{l}{H} \leq \mathbb{P}(a_h < Y \leq a_{h+1}) \leq \frac{K}{H}$ there exist two constants $0 \leq \kappa(l, K, M) < 1$,*

$C(l, K, M) > 0$ *such that for all* $j \in S_\beta$*:*

$$\sum_{h=1}^{H} \mathrm{Var}[m_j(Y) | a_h < Y \leq a_{h+1}] \leq C(l, K, M) H^{\kappa(l,K,M)} \mathrm{Var}[m_j(Y)]. \tag{3.2.1}$$

The sliced stability assumption, is an implicit assumption on the function f and the error distribution ε. If $\kappa = 0$, the condition means that the cumulative relative variability of the inverse regression curve is bounded for all slicing schemes with sufficiently small slices. If $\kappa > 0$ the cumulative relative variability of the inverse regression curve is allowed to scale sub-linearly with the number of slices.

Remark 3.2.2. *It should be expected that the sliced stability is a mild assumption. Notice that if* κ *was allowed to be* 1*, then (3.2.1) is trivially satisfied with* $C = 1/l$ *for any* H *large enough, since we have:*

$$\begin{aligned}\sum_{h=1}^{H} \mathrm{Var}[m_j(Y) | a_h < Y \leq a_{h+1}] &\leq \sum_{h=1}^{H} \mathbb{E}[m_j^2(Y) | a_h < Y \leq a_{h+1}] \\ &\leq \frac{\mathrm{Var}[m_j(Y)]}{\min_h \mathbb{P}(a_h < Y \leq a_{h+1})}.\end{aligned}$$

Finally note that $\mathbb{P}(a_h < Y \leq a_{h+1}) \geq \frac{l}{H}$*. Hence, sliced stability requires a little more than the trivial bound above to be satisfied so that the inequality will hold with an exponent* $\kappa < 1$*.*

We further assume, that the variance of the inverse regression curve $\mathrm{Var}[m_j(Y)] \propto \frac{1}{s}$, for $j \in S_\beta$, or more concretely, for some $C_V > 0$:

$$\mathrm{Var}[m_j(Y)] = \frac{C_V}{s} \text{ for } j \in S_\beta. \tag{3.2.2}$$

This assumption was originally inspired by the linear model (i.e. $Y = \beta^\intercal X + \varepsilon$), but as we show in Section 3.2.1, it turns to be a generic assumption holding for a broad class of models. Note that

$\mathrm{Var}[m_j(Y)] = 0$ for $j \notin S_\beta$, as in that case X^j is independent of Y.

The last observation, provides further the key intuition behind the method that we propose to investigate next — Diagonal Thresholding (DT). DT was suggested first by Johnstone and Lu [37], and further studied by Amini and Wainwright [3] in the sparse PCA setting. We motivate the study of DT in the SIR setting, through the fact that under the assumption (3.2.2) there is a gap between the theoretical values of the variances, and we should be able filter out the non-informative predictors by sorting out the variances $\mathrm{Var}[m_j(Y)]$.

The DT algorithm can be formulated easily along the following lines:

Algorithm 2 DT algorithm for SIR

Input: $(Y_i, X_i)_{i=1}^n$: data, H: number of slices, s: the sparsity of β

1. Calcluate $V^{jj}, j = 1, \ldots, p$ – according to formula (3.1.2);
2. Collect the s highest V^{jj} into the set $\widehat{S}$;
3. Output the set $\{j : V^{jj} \in \widehat{S}\}$.

Of course the above procedure is dependent on knowing the sparsity of the β vector. Therefore it is not realistic to use this algorithm in practical settings. Hard-thresholding can be more useful in practice, and as we show later thresholding with values in the range $[\frac{C_V}{3s}, \frac{C_V}{2s}]$ will work with high probability, provided that n is large enough. Next, we provide a theorem for the DT in the SIR framework.

Theorem 3.2.3. *Let $s = O(p^{1-\delta})$ for some $\delta > 0$. Assume that the distribution of Y is continuous, that the pair (f, ε) is sliced stable (3.2.1) holds with constants (C, l, K, M, κ) and the variance*

condition (3.2.2) with a constant C_V. Given that:

$$n \geq \Omega s \log(p-s), \tag{3.2.3}$$

for a suitably large $\Omega(C, l, K, M, \kappa, C_V)$ depending solely on constants from the sliced stability and variance assumptions, the support is recoverable by DT algorithm with probability converging to 1. *Furthermore, the number of slices H, can be held fixed (again depending on C, l, K, M, κ, C_V).*

Corollary 3.2.4. *Let $X \sim N(\mu, \mathbb{I}_{p\times p})$. Construct the "classical" SIR estimates $\widehat{V}^{jj} = \frac{1}{H}\sum_{h=1}^{H}(\overline{X}_h^j - \overline{X})^2, j = 1, \ldots, p$. Under the same assumptions as in Theorem 3.2.3, it suffices for (3.2.3) to hold in order to recover the support using DT, and in addition the number of slices can also be held fixed.*

The proof of Corollary 3.2.4 can be found in the appendix, and is a simple consequence of Theorem 3.2.3. Clearly the DT algorithm does not recover the signed support of β standalone. One can imagine applying the SIR procedure e.g. and taking the sign of the principle eigenvector, after applying DT in order to recover the signed support of β.

This motivates us to explore a procedure that has been suggested for the sparse PCA case in [18] and studied in detail by Amini and Wainwright [3]. The idea comes from the following well known characterization of the eigenvalues of a symmetric positive definite matrix:

$$\lambda_{\max}(A) = \max_{z\in\mathbb{R}:||z||_2=1} z^T A z.$$

Since we would like to require the principal eigenvector to be sparse, it would be meaningful to add the additional constraint $||z||_0 \leq s$. However this would be computationally prohibitive, and hence a reasonable formulation would be to replace the constraint with the following relaxation:

$$\operatorname*{argmax}_{z\in\mathbb{R}:||z||_2=1} z^T A z - \lambda_n ||z||_1.$$

However the above problem is maximizing a non-concave function, and thus optimization is daunting. Therefore a relaxation was suggested by d'Aspremont et al. [18], solving the following:

$$\widehat{Z} = \underset{\mathrm{tr}(Z)=1, Z\in\mathbb{S}_+^p}{\operatorname{argmax}} \mathrm{tr}(AZ) - \lambda_n \sum_{i,j=1}^{p} |Z_{ij}|. \tag{3.2.4}$$

Since this is a semidefinite program (as in this case we are looking for a maximum in the cone of the positive semidefinite matrices[3]), we refer to the optimization approach above as SDP, in future references. If the solution to the aforementioned program happens to be a rank 1 solution, then it is of the form $\widehat{Z} = \widehat{z}\widehat{z}^T$, and thus we can easily obtain an estimate of the principal eigenvector.

We summarize the signed support recovery, SDP algorithm (3.2.4) in terms of the SIR framework below.

Algorithm 3 SDP algorithm for SIR

Input: $(Y_i, X_i)_{i=1}^n$: data, H: number of slices, s: the sparsity of β

1. Calcluate the matrix V – according to formula (3.1.2);
2. Obtain the matrix $\widehat{Z}$ by solving (3.2.4), with $A = V$;
3. Find the principle eigenvector $\widehat{z}$ of $\widehat{Z}$;
4. Output $\mathrm{sign}(\widehat{z})$.

We note that this algorithm, recovers the signed support, up to multiplication by ± 1. We study the above algorithm in the SIR case, in the regime $\log s = o(\log p)$ below:

Theorem 3.2.5. *Let* $\log s = o(\log p)$. *Assume further, that the distribution of* Y *is continuous, that the pair* (f, ε) *is sliced stable (3.2.1) with constants as in Theorem 3.2.3. Then there exist a value of the tuning parameter* $\lambda_n \asymp \frac{1}{s}$ *so that Algorithm 3 recovers the signed support with probability*

converging to 1*, i.e.* $\mathbb{P}(\operatorname{sign}(\widehat{z}) = \operatorname{sign}(\beta)) \to 1$*, when:*

$$n \geq \Omega s \log(p - s), \tag{3.2.5}$$

for a large enough constant $\Omega(C, l, K, M, \kappa, C_V)$*.*

Corollary 3.2.6. *Assume that* $X \sim N(\mu, \mathbb{I})$*. Construct the "classical" SIR estimate* $\widehat{V}$ *with* $\widehat{V}^{jk} = \frac{1}{H}\sum_{h=1}^{H}(\overline{X}_h^j - \overline{X}^j)(\overline{X}_h^k - \overline{X}^k)$*. Apply the SDP algorithm to* $\widehat{V}$ *to obtain an estimate* $\widehat{\widehat{z}}$*. Then under the assumptions of Theorem 3.2.5 if (3.2.5) holds for a large enough* Ω*, we have* $\mathbb{P}(\operatorname{sign}(\widehat{\widehat{z}}) = \operatorname{sign}(\beta)) \to 1$*.*

We note that unlike Theorem 3.2.3, in Theorem 3.2.5 the number of slices H cannot be held fixed, and has to diverge slowly with $p \to \infty$. The proof of Corollary 3.2.6 can be found in the appendix.

Finally, we provide a lower bound on the sample size, under which support detection is not possible. We summarize our findings in the result below.

Theorem 3.2.7. *Let the variance condition (3.2.2) and the pair* (f, ε) *is sliced stable with constants as in Theorem 3.2.3. Then if*

$$n < \frac{1 - C_V}{C_V} 2s \log(p - s + 1),$$

the probability of any algorithm making an error on the support recovery is at least $\frac{1}{2}$ *asymptotically.*

This theorem can be seen as a converse to the previous two theorems, showing that the two algorithms are in fact achieving support recovery with an optimal sample size, up to a constant factor.

We conclude this section, by contrasting our results to the results in the sparse PCA setting, observed in Amini and Wainwright [3]. Theorem 3.2.3 presented here, shows that DT algorithm achieves an optimal rate in some sense in the SIR setting in contrast to the result in Amini and Wainwright [3]. Furthermore, Theorem 3.2.5 shows that the SDP algorithm in SIR setting can deal with a slightly

more general regime than $s = O(\log(p))$, and more importantly, it does not rely on the rank one condition, which was shown to not hold in the regime $s \gtrsim \frac{n}{\log(p)}$ by Krauthgamer et al. [41].

3.2.1 Analyzing assumptions (3.2.1) and (3.2.2)

In this section we provide sufficient conditions which ensure that (3.2.1) and (3.2.2) hold.

Variance Condition (3.2.2)

In this section we consider, a generic class of pairs of functions and errors (f, ε) satisfying the variance condition (3.2.2). Consider the following:

Lemma 3.2.8. *Let $Z \sim N(0,1)$. Let $\mathcal{F}_A = \{(f,\varepsilon) : \mathrm{Var}(\mathbb{E}[Z|f(Z,\varepsilon)]) \geq A\}$, be a subset of all pairs (f,ε) such that $f : \mathbb{R}^2 \mapsto \mathbb{R}$ and $\varepsilon \in \mathbb{R}$ be any random variable, where $0 < A \leq 1$. If $(f,\varepsilon) \in \mathcal{F}_A$, and $Y = f(\beta^\intercal X, \varepsilon)$ (where $\beta_i = \pm\frac{1}{\sqrt{s}}$ for $i \in S_\beta$ and 0 otherwise), then $\frac{A}{s} \leq \mathrm{Var}(m_j(Y)) \leq \frac{1}{s}$ for $j \in S_\beta$, where $s = |S_\beta|$.*

Proof of Lemma 3.2.8. First note that by symmetry $\beta_i m_i(Y) = \beta_j m_j(Y)$ for any $i, j \in S_\beta$. Next observe that:

$$\mathrm{Var}(\mathbb{E}[\beta^\intercal X|Y]) = \sum_{i,j\in S_\beta} \mathbb{E}[\beta_i\beta_j m_i(Y)m_j(Y)] = \sum_{i\in S_\beta} \mathrm{Var}(m_i(Y)). \tag{3.2.6}$$

Combining the observation above with the following two inequalities:

$$A \leq \mathrm{Var}(\mathbb{E}[\beta^\intercal X|Y]) \leq \mathrm{Var}(\beta^\intercal X) = 1,$$

gives the desired result. □

Remark 3.2.9. *As we can see from equation (3.2.6), if assumption (3.2.2) holds then* $\mathrm{Var}(\mathbb{E}[\beta^\intercal X|Y]) = C_V$. *This fact implies that it is necessary for a function* f *satisfying (3.2.2) to belong to all classes* $\mathcal{F}_A$, *with* $0 < A \leq C_V$.

We can thus see that the condition (3.2.2) is mild, as it simply requires the random variable $\mathbb{E}[Z|f(Z,\varepsilon)]$ to not be a constant. It is clearly implied if for example $\mathbb{E}[Zf(Z,\varepsilon)] \neq 0$. To see this assume the contrary, i.e. $\mathbb{E}[Z|f(Z,\varepsilon)] = 0$ a.s. but $\mathbb{E}[Zf(Z,\varepsilon)] \neq 0$. Then we have $\mathbb{E}[Zf(Z,\varepsilon)] = \mathbb{E}[\mathbb{E}[Z|f(Z,\varepsilon)]f(Z,\varepsilon)] = 0$, which is a contradiction.

To present the above abstract framework in a more illustrative fashion, we consider several examples. We start with the following simple model:

$$Y = f(\beta^\intercal X + \varepsilon) \text{ with } \varepsilon \sim N(0,\sigma^2), \tag{3.2.7}$$

where f is a univariate, continuous monotone function. Let $Z = \beta^\intercal X \sim N(0,1)$. Looking into:

$$\mathbb{E}[Z|f(Z+\varepsilon) = c] = \mathbb{E}[Z|Z+\varepsilon = f^{-1}(c)] = \frac{f^{-1}(c)}{1+\sigma^2}, \tag{3.2.8}$$

where the last equality follows from the multivariate normal distribution properties. Thus:

$$\mathrm{Var}(\mathbb{E}[Z|f(Z+\varepsilon)]) = \frac{1}{(1+\sigma^2)^2}\,\mathrm{Var}(Z+\varepsilon) = \frac{1}{1+\sigma^2}.$$

To see a slightly different example, consider a setting which is a special case of the single index model: $Y = f(\beta^\intercal X) + \varepsilon$, where $f : \mathbb{R} \mapsto \mathbb{R}$ is a continuous and increasing function with $\inf_{z\in\mathbb{R}} f(z) = -\infty$ and $\sup_{z\in\mathbb{R}} f(z) = +\infty$, and the random variable $|\varepsilon|$ has a bounded support by $M > 0$. In this setting it is clear that the random variable $\mathbb{E}[Z|f(Z)+\varepsilon]$ is non-constant as

$$f^{-1}(c-M) \leq \mathbb{E}[Z|f(Z)+\varepsilon = c] \leq f^{-1}(c+M).$$

Therefore if $c < -M$ and $c' > M$ it follows that $\mathbb{E}[Z|f(Z)+\varepsilon = c] < \mathbb{E}[Z|f(Z)+\varepsilon = c']$. Of course, clearly the condition in this example is far from being necessary.

As a counter-example to the variance condition consider $f(Z,\varepsilon) = g(Z)$, where g is some even function. Then we have $\mathbb{E}[Z|g(Z)] = 0$, and thus we can't claim that the variance will be of order $\frac{1}{s}$. In fact it is clear that detection based on conditional variance in this case is impossible, as (3.2.6) in fact gives us that $\mathrm{Var}[m_j(Y)] = 0$ for all j, and therefore the variance of the sliced inverse regression curve, contains no information on the support of β.

Sliced Stability (3.2.1)

We start this section by showing that the simple example (3.2.7) considered in the section 3.2.1 satisfies (3.2.1). Using (3.2.8) we immediately get:

$$\frac{1}{s}\sum_h \mathrm{Var}\left[\frac{f^{-1}(Y)}{1+\sigma^2}\middle| a_h < Y \le a_{h+1}\right] = \frac{1}{(1+\sigma^2)s}\sum_h \mathrm{Var}\left[\frac{V}{\sqrt{1+\sigma^2}}\middle| f^{-1}(a_h) < V \le f^{-1}(a_{h+1})\right],$$

where $V = f^{-1}(Y) \sim N(0, 1+\sigma^2)$. It is clear that out of all sets of probability q the set giving the maximal variance for the truncated normal distribution is the one symmetric about 0. Therefore, using the truncated normal distribution properties, we can bound the expression above by:

$$\frac{1}{(1+\sigma^2)s} H\left(1 - \frac{2\Phi^{-1}(\frac{1}{2}+\frac{q}{2})\phi(\Phi^{-1}(\frac{1}{2}+\frac{q}{2}))}{q}\right),$$

where $q = \max_h \mathbb{P}\left(\frac{f^{-1}(a_h)}{\sqrt{1+\sigma^2}} < Z \le \frac{f^{-1}(a_{h+1})}{\sqrt{1+\sigma^2}}\right)$, with $Z \sim N(0,1)$. In the appendix we show the following:

Lemma 3.2.10. *For $q \in [0, 2\Phi(r)-1)$, for some $r > 0$ we have the following bound:*

$$\left(1 - \frac{2\Phi^{-1}(\frac{1}{2}+\frac{q}{2})\phi(\Phi^{-1}(\frac{1}{2}+\frac{q}{2}))}{q}\right) \le \frac{q}{8}\frac{r}{\phi(r)}.$$

Therefore if $q \leq \frac{K}{H}$ we have by Lemma 3.2.10 that, for $H \geq M = \frac{K}{2\Phi(r)-1}$, for some $r > 0$:

$$\frac{1}{s}\sum_{h} \mathrm{Var}\left[\frac{f^{-1}(Y)}{1+\sigma^2}\middle| a_h < Y \leq a_{h+1}\right] \leq \frac{K}{(1+\sigma^2)8}\frac{r}{\phi(r)}\frac{1}{s}.$$

In other words this model satisfies the sliced stability condition with $\kappa(l, K, M) = 0, C(l, K, M) = \frac{K}{(1+\sigma^2)8}\frac{r}{\phi(r)}/C_V$. Note furthermore, that in this special case sliced stability holds with $l = 0$.

Next, we proceed to formulate a more generic sufficient condition implying sliced stability. We borrow ideas from Hsing and Carroll [33], and show that their well accepted sufficient conditions imply sliced stability, with a slight modification (see Remark 3.2.12).

Let $\mathcal{A}_H(l, K)$, with $K > 1, 0 < l < 1$, denote all partitions of $\mathbb{R}$ of the sort $\{-\infty = a_1 \leq a_2 \leq \ldots \leq a_{H+1} = +\infty\}$, such that $\frac{l}{H} \leq \mathbb{P}(a_h \leq Y \leq a_{h+1}) \leq \frac{K}{H}$.

Moreover, for any fixed $B \in \mathbb{R}$, let $\Pi_r(B)$ denote all possible partitions of the closed interval $[-B, B]$ into r points $-B \leq b_1 \leq b_2 \leq \ldots \leq b_r \leq B$.

Define the normalized version of the centered inverse regression curve $m(y) = \frac{m_j(y)}{\sqrt{\mathrm{Var}(m_j(Y))}}$, and let m satisfy the following smoothness condition:

$$\lim_{r\to\infty} \sup_{b\in\Pi_r(B)} r^{-1/(2+\xi)} \sum_{i=2}^{r} |m(b_i) - m(b_{i-1})| = 0, \tag{3.2.9}$$

for any $B > 0$ for some fixed $\xi > 0$. Note that as mentioned in Hsing and Carroll [33], assumption (3.2.9) is weaker than the assumption that m is of bounded variation, and furthermore the bigger the ξ the more stringent this assumption becomes. In addition assume that, there exists $B_0 > 0$ and a non-decreasing function $\widetilde{m} : (B_0, \infty) \mapsto \mathbb{R}$, such that:

$$|m(x) - m(y)| \leq |\widetilde{m}(|x|) - \widetilde{m}(|y|)|, \text{ for } x, y \in (-\infty, B_0) \text{ or } (B_0, \infty), \tag{3.2.10}$$

and moreover, $\mathbb{E}[|\widetilde{m}(|Y|)|^{(2+\xi)}] < \infty$ (where in the expectation we set $\widetilde{m}(y) = 0$ for $|y| \leq B_0$).

Remark 3.2.11. *We note that without loss of generality we can consider the function $\widetilde{m}$ to be non-negative, at the price of potentially shrinking the interval (B_0, ∞) to $(B_0 + \epsilon, \infty)$ by any $\epsilon > 0$. To see this fix an $\epsilon > 0$, and define $\widetilde{m}'(x) = m(x) - m(B_0 + \epsilon)$ for $x \in (B_0 + \epsilon, \infty)$. Then since (3.2.10) holds on $(-\infty, -B_0) \cup (B_0, \infty)$, clearly:*

$$|m(x) - m(y)| \leq |\widetilde{m}'(|x|) - \widetilde{m}'(|y|)|, \text{ for } x, y \in (-\infty, -B_0 - \epsilon) \text{ or } (B_0 + \epsilon, +\infty).$$

By the convexity of the map $x \mapsto x^{2+\xi}$ we have $\widetilde{m}'(x)^{2+\xi} \leq 2^{1+\xi}(\widetilde{m}(x)^{2+\xi} + \widetilde{m}(B_0 + \epsilon)^{2+\xi})$ and hence $\mathbb{E}[|\widetilde{m}'(|Y|)|^{(2+\xi)}] < \infty$. Finally by definition $\widetilde{m}'$ is non-negative and non-decreasing on $(B_0 + \epsilon, \infty)$. From now on we will consider $\widetilde{m}$ to be non-negative on (B_0, ∞) without further reference.

Remark 3.2.12. *The moment inequality implies a tail condition, i.e. $|\widetilde{m}(y)|^{(2+\xi)}\mathbb{P}(|Y| > y) \to 0$. A tail condition is assumed in[33] of the sort $\widetilde{m}^4(y)\mathbb{P}(|Y| > y) \to 0$ when $y \to \infty$[†]. We note that this tail condition is just slightly weaker than assuming that $\mathbb{E}[\widetilde{m}^4(|Y|)] = \int_0^\infty \widetilde{m}^4(y)d\mathbb{P}(|Y| \leq y) < \infty$. To see this, from the previous equation it's clear that $\mathbb{E}[\widetilde{m}^4(|Y|)] < \infty$ implies $\widetilde{m}^4(y)P(|Y| > y) \to 0$ since $\int_z^\infty \widetilde{m}^4(y)d\mathbb{P}(|Y| \leq y) \geq m^4(z)\mathbb{P}(|Y| > z)$ for any z. On the other hand, the fact that $\widetilde{m}^4(y)\mathbb{P}(|Y| > y) \to 0$ implies $u^4\mathbb{P}(\widetilde{m}(|Y|) > u) \to 0$, as $u \to \widetilde{m}(+\infty)$ (since $\mathbb{P}(\widetilde{m}(|Y|) > \widetilde{m}(y)) \leq \mathbb{P}(|Y| > y)$). Using the representation:*

$$\begin{aligned}\mathbb{E}\widetilde{m}^{4-\epsilon}(|Y|) &= \int_0^\infty (4-\epsilon)u^{3-\epsilon}\mathbb{P}(\widetilde{m}(|Y|) > u)du \\ &\leq (4-\epsilon) + \int_1^\infty (4-\epsilon)\frac{u^4\mathbb{P}(\widetilde{m}(|Y|) > u)}{u^{1+\epsilon}}du < \infty.\end{aligned}$$

Hence, we conclude that $\mathbb{E}\widetilde{m}^{4-\epsilon}(|Y|) < \infty$ for any small $\epsilon > 0$.

[†]The degree 4 was selected in[33], because they were seeking $\sqrt{n}$ consistency, which is not required in our setting. Hence our condition requires a less stringent degree.

We are now in a position to formulate the following:

Proposition 3.2.13. *Assume that the standardized centered inverse regression curve satisfies properties (3.2.9) and (3.2.10) for some $\xi > 0$. Then we have that for any fixed $0 < l < 1 < K$:*

$$\lim_{H\to\infty} \sup_{a\in\mathcal{A}_H(l,K)} \frac{1}{H^{2/(2+\xi)}} \sum_{h=1}^{H} \mathrm{Var}[m(Y)|a_h < Y \leq a_{h+1}] \to 0. \tag{3.2.11}$$

We defer the proof of Proposition 3.2.13 to the Appendix. It is clear however that (3.2.11) implies the existence of constants $M, C(l, K, M)$ such that (3.2.1) holds, with $\kappa = \frac{2}{2+\xi} < 1$.

We conclude this section, by recalling several remarks mentioned in Hsing and Carroll [33], regarding the mildness of their conditions, which as we saw imply sliced stability. Firstly, condition (3.2.9) is weaker than requiring that m has bounded variation. Secondly, in the case when m is a continuous and increasing function, we can select $\widetilde{m}(x) = |m(|x|)|$, and provided that $\mathbb{E}|m(|Y|)|^{2+\xi} < \infty$ this readily implies both (3.2.9) and (3.2.10).

3.3 Numerical Results

In this section we consider several models to evaluate the predictions of Theorems 3.2.3 and 3.2.5 numerically. We consider the following scenarios:

$$Y = \sin(\beta^{\intercal} X) + U(0, 1), \tag{3.3.1}$$

$$Y = (\beta^{\intercal} X)^3 + N(0, 1), \tag{3.3.2}$$

$$Y = (\beta^{\intercal} X + N(0, 1))^3, \tag{3.3.3}$$

$$Y = \beta^{\intercal} X + N(0, 1). \tag{3.3.4}$$

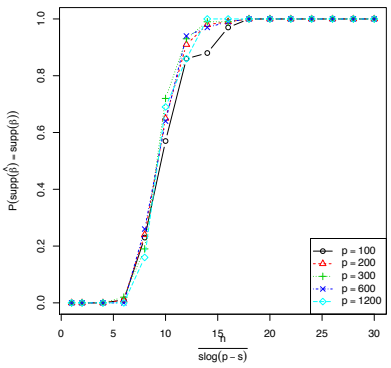
Note that these models do not necessarily fall into the same class of sliced stability and variance

assumptions. Therefore we would not expect to see the phase transition, described in Theorems 3.2.3 and 3.2.5, to occur at the same places for all models.

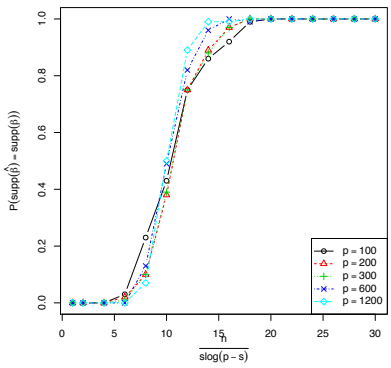
We first explore the predictions of Theorem 3.2.3. Even though we provide theoretical values of the constants H and m, we ran all simulations with $H = 10$ slices. We believe this scenario, is still reflective of the true nature of the DT algorithm, as the theoretical value of H we provide is not optimized in any fashion.

In figure 3.1, we present DT results from plots for different p values in the regime $s = \sqrt{p}$.

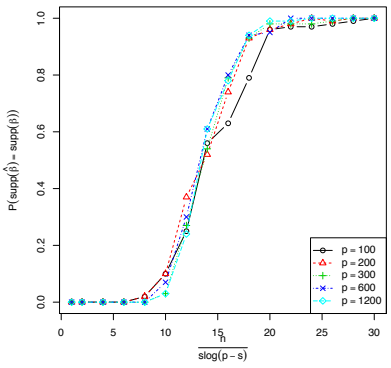
Figure 3.1: DT, $s = \sqrt{p}$



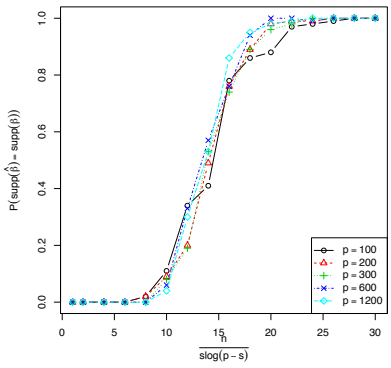
(a) Model (3.3.1)



(b) Model (3.3.2)



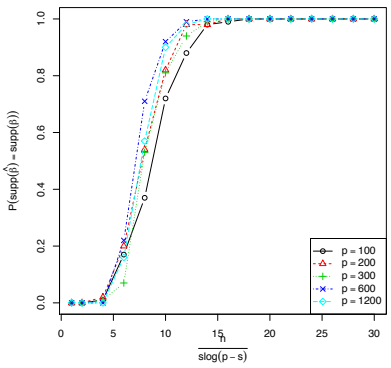
(c) Model (3.3.3)



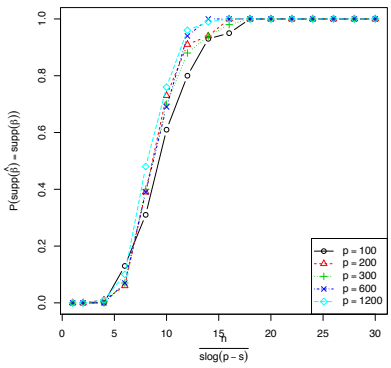
(d) Model (3.3.4)

The plots in the case $s = \sqrt{p}$, are really similar to the corresponding plots in the regime $s = \log(p)$, which can be seen in figure 3.2:

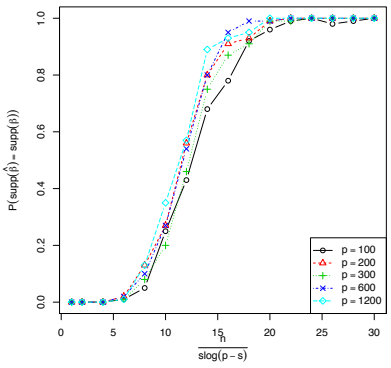
Figure 3.2: DT, $s = \log(p)$



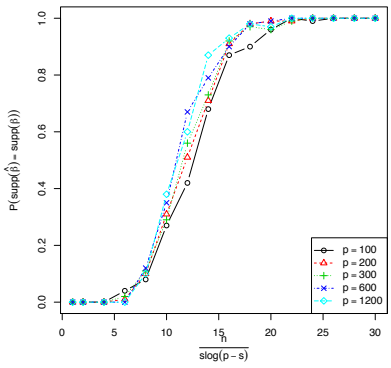
(a) Model (3.3.1)



(b) Model (3.3.2)



(c) Model (3.3.3)



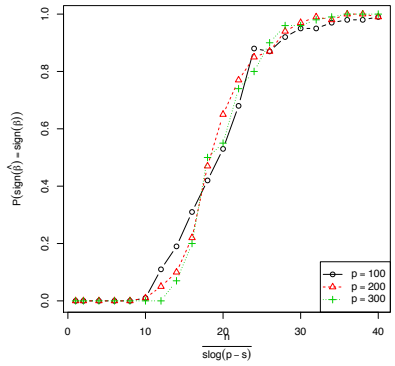
(d) Model (3.3.4)

This similarity is in concordance with the predictions from our theoretical results. We can distinctly see the phase transition occurring in approximately the same place regardless of the values of the dimension p, that we use.

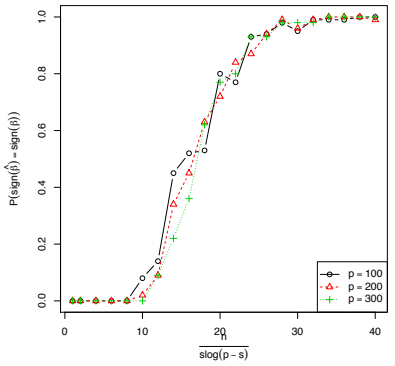
Finally, we present the corresponding results for algorithm 3. We used the code from an efficient implementation of the program (3.2.4), as suggested by Zhang and Ghaoui [97]. The code was kindly provided to us by the authors of [97]. In figure 3.3 we provide the four models for the case when $s =$

$\log(p)$ (and hence $\log s = o(\log(p))$).

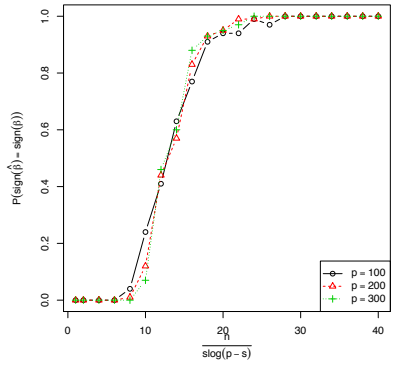
Figure 3.3: SDP, $s = \log(p)$



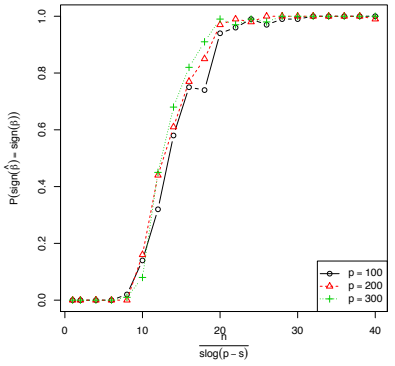
(a) Model (3.3.1)



(b) Model (3.3.2)



(c) Model (3.3.3)



(d) Model (3.3.4)

Here we have again used $H = 10$ in all scenarios, for simplicity. We observe that phase transitions are occurring in all of the models with the signed support being correctly recovered for large enough values of $\frac{n}{s\log(p-s)}$. We note that the phase transition for SDP does not seem to occur at the same values as the phase transition for DT. Observe that our results do not contradict this fact.

3.4 Proof of Theorem 3.2.3

We first present the high level outline of the proof. Note that for vectors with $j \notin S_\beta$, we have that X_i^j are completely independent of the slicing scheme on Y and therefore the element $V^{jj} \sim \frac{1}{mH}\chi^2_H$.

We would thus expect, to be able to filter out unrelated variables, by selecting the highest values of the diagonal elements. Our argument shows that in fact $V^{jj} \geq \frac{1}{2}\operatorname{Var}[m_j(Y)]$, for all $j \in S_\beta$ with high probability. To achieve this we would like to control the quantity:

$$|V^{jj} - \operatorname{Var}[m_j(Y)]| = \left| \frac{1}{H}\sum_{h=1}^{H} \left(\overline{X}_h^j\right)^2 - \int m_j^2(y) p_Y(y) dy \right|. \tag{3.4.1}$$

We will show below (see (3.4.5)), that the above expression is well approximated by:

$$|V^{jj} - \operatorname{Var}[m_j(Y)]| \approx \left| \frac{1}{H}\sum_{h=1}^{H} \left(\overline{X}_h^j\right)^2 - \sum_{h=1}^{H} (\mu_h^j)^2 \mathbb{P}(Y \in S_h) \right|, \tag{3.4.2}$$

under sliced stability (where $Y_{(0)} = -\infty$). Our proof then controls (3.4.2), by rigorously exploring the intuitive facts that $\mathbb{P}(Y \in S_h) \approx \frac{1}{H}$ and $\overline{X}_h^j \approx \mu_h^j$. To deal with the former approximation we use the Dvoretzky-Kiefer-Wolfowitz (DKW) inequality (see Massart [61]), and we develop a new concentration inequality based on Bernstein's inequality to deal with the latter one. We now proceed to rigorously show the result.

Note that the probability $\mathbb{P}(Y \in S_h)$ is a random variable, where the randomness comes from the two ends $Y_{(m(h-1))}$ and $Y_{(mh)}$ of the interval S_h. Recall that F_n is the empirical distribution function of Y, based on the sample Y_i. Since we are assuming Y is coming from a continuous distri-

bution, we have:

$$\begin{aligned}\mathbb{P}(Y \in S_h) &\leq \mathbb{P}(F_n(Y_{(m(h-1))}) \leq F_n(Y) \leq F_n(Y_{(mh)})) \\ &= \mathbb{P}\left(\frac{h-1}{H} \leq F_n(Y) \leq \frac{h}{H}\right),\end{aligned}$$

where with abuse of notation ($Y_{(0)} = -\infty$). Conversely we also have:

$$\begin{aligned}\mathbb{P}(Y \in S_h) &\geq \mathbb{P}(F_n(Y_{(m(h-1))}) < F_n(Y) < F_n(Y_{(mh)})) \\ &= \mathbb{P}\left(\frac{h-1}{H} < F_n(Y) < \frac{h}{H}\right).\end{aligned}$$

Now using the DKW inequality, we have that $\mathbb{P}(\sup_Y |F_n(Y) - F(Y)| > \epsilon) \leq 2\exp(-2n\epsilon^2)$, which in conjunction with the fact that Y comes from a continuous distribution, implies that for all h we have:

$$\frac{1}{H} - 2\epsilon \leq \mathbb{P}\left(\frac{h-1}{H} < F_n(Y) < \frac{h}{H}\right) \leq \mathbb{P}\left(\frac{h-1}{H} \leq F_n(Y) \leq \frac{h}{H}\right) \leq \frac{1}{H} + 2\epsilon, \quad (3.4.3)$$

on an event with probability at least $1 - 2\exp(-2n\epsilon^2)$. Let the event where this bound holds is S.

We now describe two different ways that we can use to generate a data from the SIR model. The straightforward way to generate data from the SIR model (3.1.1) is the "forward" way — by first generating $X \sim N(0, \mathbb{I})$, next independently generating some random noise ε from its corresponding distribution, and finally generating a $Y = f(\beta^\intercal X, \varepsilon)$. In doing so notice that the Y values will be generated in no particular order. In the second approach, we describe a two-step generation procedure. In the first step consider generating values of Y coming from the joint distribution of the order statistics — $(Y_{(m)}, Y_{(2m)}, \ldots, Y_{((H-1)m)})$. We can then, conditionally on the $(Y_{(m)}, Y_{(2m)}, \ldots, Y_{((H-1)m)})$ values, generate corresponding predictors $(X_{(m)}, X_{(2m)}, \ldots, X_{((H-1)m)})$

and random noise $(\varepsilon_{(m)}, \varepsilon_{(2m)}, \ldots, \varepsilon_{((H-1)m)})$ independently, with each pair $(X_{(mh)}, \varepsilon_{(mh)})$ coming from the conditional joint distribution $(X, \varepsilon)|f(\beta^\intercal X, \varepsilon) = Y_{(mh)}$. In the second step, for each interval S_h, we can use rejection sampling by generating $X \sim N(0, \mathbb{I})$ and independently random noise ε and accepting X iff $Y_{(m(h-1))} < f(\beta^\intercal X, \varepsilon) \leq Y_{(mh)}, 1 \leq h < H-1$ and $Y_{(m(H-1))} < f(\beta^\intercal X, \varepsilon)$ for $h = H-1$. We will do so for each of the intervals until we accept $m-1$ points in the first $H-1$ intervals and m points in the last one. Once we have the X and ε values it's straightforward to calculate the remaining Y values.

The second data generation mechanism which we described above, gives us the insight that conditionally on the values $(Y_{(m)}, Y_{(2m)}, \ldots, Y_{((H-1)m)})$ the sample means $\overline{X}^j_{h,1:(m-1)}$ have corresponding population mean — μ^j_h for $h = 1, \ldots, H-1$ and the sample mean $\overline{X}^j_H$ has a mean of μ^j_H. To be consistent with the notations of the other slices, let us randomly select a point in the H^{th} slice and discard it from the means. With a slight abuse of notation we will denote the average of the remaining points in the H^{th} slice $\overline{X}^j_{H,1:(m-1)}$ and the discarded point $X_{H,m}$ keeping in mind that this point need not be the m^{th} point, but was chosen arbitrarily, so that the mean is still equal to μ^j_H. We next formulate the following key concentration result for the sliced means, which we show in the appendix:

Lemma 3.4.1. *On the event S, for $\eta > 0$ we have the following:*

$$\mathbb{P}\left(\max_{j \in S_\beta, h \in \{1,\ldots,H\}} \left|\overline{X}^j_{h,1:(m-1)} - \mu^j_h\right| > \eta\right) \leq$$
$$2sH \exp\left(-\frac{1}{2}\frac{\eta^2(m-1)}{\eta + \widetilde{C}_1 + \widetilde{C}_2\sqrt{-\log\left(\frac{q}{2}\right)} + \widetilde{C}_3\left(-\log\left(\frac{q}{2}\right)q\right)}\right), \tag{3.4.4}$$

with $\widetilde{C}_i, i = 1, 2, 3$ being fixed constants, and $q = \frac{1}{H} - 2\epsilon$ (assuming that H is sufficiently large so that $q < 2 - 2\Phi(1/\sqrt{\sqrt{2}-1})$).

Denote with $\widetilde{S}$ the event on which we have

$$\max_{j\in S_\beta, h\in\{1,\dots,H\}} \left|\overline{X}^j_{h,1:(m-1)} - \mu^j_h\right| \leq \eta.$$

By (3.4.4), (3.4.3) and the union bound we have that:

$$\mathbb{P}(\widetilde{S}) \geq 1 - 2sH\exp\left(-\frac{1}{2}\frac{\eta^2(m-1)}{\eta + \widetilde{C}_1 + \widetilde{C}_2\sqrt{-\log\left(\frac{q}{2}\right)} + \widetilde{C}_3\left(-\log\left(\frac{q}{2}\right)q\right)}\right)$$
$$- 2\exp(-2n\epsilon^2).$$

Next we move on, to show that (3.4.1) is close to (3.4.2) on the event S, as well as we collect two straightforward inequalities in the following helpful:

Lemma 3.4.2. *Assume that the sliced stability condition (3.2.1) holds. Then we have the following inequalities holding on the event S, for large enough H, and small enough ϵ:*

$$|V^{jj} - \mathrm{Var}[m_j(Y)]| \leq \left|\frac{1}{H}\sum_{h=1}^{H}\left(\overline{X}^j_h\right)^2 - \sum_{h=1}^{H}(\mu^j_h)^2\mathbb{P}(Y\in S_h)\right| + \underbrace{\frac{CH^\kappa}{s}\left(\frac{1}{H}+2\epsilon\right)}_{B_1}, \tag{3.4.5}$$

$$\sum_{h=1}^{H}(\mu^j_h)^2 \leq \underbrace{\frac{\frac{C_V}{s} + \frac{CH^\kappa}{s}\left(\frac{1}{H}+2\epsilon\right)}{\left(\frac{1}{H}-2\epsilon\right)}}_{B_2}, \tag{3.4.6}$$

$$\sum_{h=1}^{H}|\mu^j_h| \leq \underbrace{\frac{\sqrt{\frac{C_V}{s} + \frac{CH^\kappa}{s}\left(\frac{1}{H}+2\epsilon\right)}}{\left(\frac{1}{H}-2\epsilon\right)}}_{B_3}. \tag{3.4.7}$$

Note. *We refer to the constants from (3.2.1) as C and κ, dropping the dependence on K and M for brevity, and in fact $C = C(l,K,M)C_V$.*

Note that by an elementary calculation — using (3.4.3) and Lemma 3.4.2, on the event $\widetilde{S}$ we get:

$$|V^{jj} - \mathrm{Var}[m_j(Y)]| \leq \frac{1}{H}\sum_{h=1}^{H}\left|\left(\overline{X}_h^j\right)^2 - \frac{(m-1)^2}{m^2}(\mu_h^j)^2\right| + \left(2\epsilon + \frac{1}{H} - \frac{(m-1)^2}{Hm^2}\right)B_2 + B_1, \tag{3.4.8}$$

where we used (3.4.5), the triangle inequality and (3.4.6). Consider the following:

Lemma 3.4.3. *There exists a subset $\widetilde{\widetilde{S}} \subset \widetilde{S}$ such that $\mathbb{P}(\widetilde{S} \setminus \widetilde{\widetilde{S}}) \leq s\exp(-\frac{3}{16}n\tau^2)$, for any fixed $\tau \in [0, \frac{1}{2})$ on which we have the following bound for any $j \in S_\beta$:*

$$\frac{1}{H}\sum_{h=1}^{H}\left|\left(\overline{X}_h^j\right)^2 - \frac{(m-1)^2}{m^2}(\mu_h^j)^2\right| \leq \frac{(1+\tau)}{m} + \frac{2\sqrt{1+\tau}}{\sqrt{m}}\sqrt{2\eta^2 + 2\frac{B_2}{H}} + \eta^2 + 2\eta\frac{B_3}{H}. \tag{3.4.9}$$

We defer the proof of this lemma till the appendix.

Next, we provide exact constants, such that each of the six terms in inequalities (3.4.8) and (3.4.9) bounding $|V^{jj} - \mathrm{Var}[m_j(Y)]|$ are $\leq \frac{C_V}{12s}$, and the probability of the event $\widetilde{\widetilde{S}}$ still converges to 1. The remarkable phenomenon here is that the number of slices H, can be selected so that it is a constant, which might seem counterintuitive. Select the constants in the following manner:

$$H = \max\left\{M, \left(\frac{12CK}{C_V}\right)^{\frac{1}{1-\kappa}}, \frac{K}{2}\exp(1), \frac{1}{2(1-\Phi((\sqrt{2}-1)^{-1/2}))}\right\}, \tag{3.4.10}$$

$$\epsilon = \min\left\{\frac{K-1}{2H}, \frac{1-l}{2H}, \frac{1}{54H}\right\}, \tag{3.4.11}$$

$$\eta = \frac{\widetilde{C}_0}{\sqrt{s}}, \tag{3.4.12}$$

$$m \geq 104 + \max\left\{2\frac{\widetilde{C}_4}{\widetilde{C}_0^2}(1+\gamma), \frac{\widetilde{C}_5}{C_V}\right\}s\max(\log(s+1), \log(p-s)), \tag{3.4.13}$$

where $\gamma > 0$ is any positive constant, and $\widetilde{C}_0 = \frac{1}{48\sqrt{1+\frac{1}{12}}}\sqrt{C_V}$, $\widetilde{C}_4 = \widetilde{C}_0 + \widetilde{C}_1 + \widetilde{C}_2\sqrt{-\log\frac{1}{4H}} + \widetilde{C}_3(-\log(\frac{K}{2H})\frac{K}{H})$ and $\widetilde{C}_5 = 12^2 4(1+\tau)(\frac{1}{6}+\frac{1}{3}+4)$. We show in Appendix B.2 that these constants keep $\mathbb{P}(\widetilde{\widetilde{S}}) \to 1$, and satisfy the requirement that $|V^{jj} - \mathrm{Var}[m_j(Y)]| \leq \frac{C_V}{2s}$, and therefore give us the following bound with high probability: $V^{jj} \geq \frac{C_V}{2s}$. Note that we have made no effort whatsoever to optimize any of these constants and they are just an example, which can easily be bettered.

Finally we need to deal with the variables V^{jj} for $j \notin S_\beta$. We make usage of the following tail bound, for χ^2 random variables which we take from Laurent and Massart [42] (see Lemma 1):

$$\mathbb{P}\left(\frac{\chi^2_H}{H} \geq 1 + 2\sqrt{\frac{x}{H}} + \frac{2x}{H}\right) \leq \exp(-x).$$

Note that, $V^{jj} \sim \frac{1}{mH}\chi^2_H$ for $j \in S^c_\beta$. Thus applying the bound above we have

$$\frac{1}{mH}\chi^2_H \leq \frac{1}{m} + \frac{2}{m}\sqrt{\frac{x}{H}} + \frac{2x}{mH}, \tag{3.4.14}$$

with probability at least $\exp(-x)$. We select x in such a manner so that we make sure $x - \log(p-s) \to +\infty$ which will guarantee by the union bound that all bounds will hold with high probability. Moreover, we require each of the three terms on the RHS of (3.4.14) to be bounded by $\frac{C_V}{12s}$, which will ensure that each of the V^{jj} for $j \in S^c_\beta$ will be bounded from above by $\frac{C_V}{4s}$ and we will be able to threshold by choosing a cutoff of $\frac{C_V}{3s}$. By choosing $x = \frac{C_V}{24}\frac{n}{s}$, we can ensure that all three terms will satisfy the requirement we imposed above, since (3.4.13) gives $x = \frac{C_V}{24}\frac{n}{sH} \geq 1$ and thus the maximum term of the three is $\frac{2x}{mH} \leq \frac{C_V}{12s}$. Finally if we have:

$$\frac{C_V}{24}\frac{n}{s} > 2\log(p-s),$$

it will follow that with probability tending to 1 we can separate the signals. We conclude that if:

$$n > \max\left(Hm, \frac{48}{C_V} s \log(p-s)\right),$$

detection is possible. Of course the last inequality is $n > \Omega s \log(p-s)$ asymptotically in the regime $s = O(p^{1-\delta})$, where $\Omega = \max(H\widetilde{\widetilde{C}}, \frac{48}{C_V})$ with $\widetilde{\widetilde{C}}$ determined through (3.4.13). This is what we wanted to show.

□

3.5 Proof of Theorem 3.2.5

In this section we show that under the assumption $\log s = o(\log p)$, the SDP relaxation will have a rank 1 solution with high probability and moreover this solution will recover the signed support of the vector β. For the analysis of the algorithm we set the regularization parameter $\lambda_n = \frac{C_V}{2s}$.

To this end, we restate Lemma 5 from Amini and Wainwright [3], which provides a sufficient condition for a global solution of the SDP problem:

Lemma 3.5.1. *Suppose there exists a matrix U satisfying:*

$$U_{ij} = \begin{cases} \operatorname{sign}(\widehat{z}_i)\operatorname{sign}(\widehat{z}_j), & \text{if } \widehat{z}_i\widehat{z}_j \neq 0; \\ \in [-1,1], & \text{otherwise.} \end{cases} \tag{3.5.1}$$

Then if $\widehat{z}$ is the principle eigenvector of the matrix $A - \lambda_n U$, $\widehat{z}\widehat{z}^\intercal$ is the optimal solution to problem (3.2.4).

Recall that the SIR estimate of the variance-covariance matrix has entries:

$$V^{jk} = \frac{1}{H}\sum_{h=1}^{H}\left(\frac{1}{m}\sum_{i=1}^{m} X_{h,i}^{j}\right)\left(\frac{1}{m}\sum_{i=1}^{m} X_{h,i}^{k}\right).$$

Denote with $\widetilde{V} = V - \lambda_n U$, where U is to be defined matrix from Lemma 3.5.1.

We furthermore consider the decomposition of $\widetilde{V}$ into blocks – $\widetilde{V}_{S_\beta,S_\beta}$, $\widetilde{V}_{S_\beta^c,S_\beta}$, $\widetilde{V}_{S_\beta^c,S_\beta^c}$. Here, these three matrices are sub matrices of the matrix V restricted to entries with indexes in the sets S_β or S_β^c correspondingly.

We first focus on the V_{S_β,S_β} matrix. We calculate the value of the covariance of two coordinates $j, k \in S_\beta$:

$$\begin{aligned}\mathrm{Cov}[m_j(Y), m_k(Y)] &= \mathbb{E}[m_j(Y), m_k(Y)] \\ &= \mathrm{sign}(\beta_j)\,\mathrm{sign}(\beta_k)\mathbb{E}[m_k^2(Y)] \\ &= \beta_j\beta_k C_V,\end{aligned} \tag{3.5.2}$$

where we used that $\mathrm{sign}(\beta_j)m_j(Y) = \mathrm{sign}(\beta_k)m_k(Y)$, which follows by noticing that the distribution of $X^j|Y$ is the same as the distribution of $X^k|Y$ except the potential difference in the signs of the coefficients, because of the symmetry in the problem.

We proceed with formulating a bound similar to Lemma 3.4.3, but for the covariance:

Lemma 3.5.2. *On the event $\widetilde{\widetilde{S}}$ as defined in Lemma 3.4.3, for $j, k \in S_\beta$ and $j \neq k$, we have the following inequality:*

$$\begin{aligned}\left|V^{jk} - \mathrm{Cov}(m_j(Y), m_k(Y))\right| \leq &\left(2\epsilon + \frac{1}{H} - \frac{(m-1)^2}{Hm^2}\right)B_2 + B_1 + 4\eta\frac{B_3}{H} \\ &+ \frac{4(1+\tau)}{m} + \frac{4\sqrt{1+\tau}}{\sqrt{m}}\sqrt{2\eta^2 + 2\frac{B_2}{H}} + 4\eta^2.\end{aligned} \tag{3.5.3}$$

Let $t = \log\left(\frac{\log(p)}{\log(s)}\right) + 1$, and note that $t \to \infty$ with $p \to \infty$ since $\log s = o(\log p)$. Choose the constants in the following manner:

$$H = \max\left\{M, \left(\frac{12CKt}{C_V}\right)^{\frac{1}{1-\kappa}}, \frac{K}{2}\exp(1), \frac{1}{2(1-\Phi((\sqrt{2}-1)^{-1/2}))}\right\}, \tag{3.5.4}$$

$$\epsilon = \min\left\{\frac{K-1}{2H}, \frac{1-l}{2H}, \frac{l}{4H(1+12t)}\right\}, \tag{3.5.5}$$

$$\eta = \frac{\sqrt{C_V}l}{4(12t+1)\sqrt{s}}, \tag{3.5.6}$$

$$\begin{aligned} m \geq \max\Bigg\{ & \frac{(1+\tau)48^2 st^2\left(\frac{1}{24}+\frac{2}{l}+\frac{1}{6l}\right)}{C_V} \\ & \frac{16s(12t+1)^2 t\log(sH)(\widetilde{C}' + \widetilde{C}_2\sqrt{\log H})}{l^2 C_V} \\ & Ht^3, \frac{4(12t+1)}{l}\Bigg\}, \end{aligned} \tag{3.5.7}$$

where the constant $\widetilde{C}'$ is defined in the supplement. It is not hard to see, using the fact that $\log s = o(\log p)$ (shown in the supplement), that using the updated constants above we can get the following bound

$$\sup_{j,k\in S_\beta}\left|V^{jk} - \mathrm{Cov}(m_j(Y), m_k(Y))\right| \leq \frac{C_V}{2st}, \tag{3.5.8}$$

on $\widetilde{\widetilde{S}}$, with the probability of $\widetilde{\widetilde{S}}$ tending to 1. The idea for the constant selection here is identical to the one in the DT case, but we required each of the 6 terms in (3.5.3) to be smaller than $\frac{C_V}{12st}$. We note that in the case of the SDP algorithm, we need to let H diverge to infinity. Furhtermore we note, that elementary calculation shows that mH is of lower order than $s\log(p)$, using the fact that $\log(s) = o(\log(p))$.

Having in mind the above inequality we consider the matrix $\widetilde{V}_{S_\beta,S_\beta}$:

$$\widetilde{V}_{S_\beta,S_\beta} = \frac{C_V}{2}\beta_{S_\beta}\beta_{S_\beta}^{\intercal} + N,$$

where N is some symmetric noise matrix. Note that by (3.5.2), (3.5.8) gives a bound on $||N||_{\max}$. We next verify that the matrix N satisfies the conditions of Lemma 6 in Amini and Wainwright [3].

Take any s-dimensional unit vector $||v||_2 = 1$, and calculate:

$$|v^{\intercal}Nv| \leq ||v||_1^2||N||_{\max} \leq \frac{C_V}{2st}s||v||_2^2 = \frac{C_V}{2t},$$

which converges to 0, as $p \to \infty$, by the definition of t. This implies that $||N||_{2,2} \to 0$ as $p \to \infty$, since $||N||_{2,2} = |\lambda_{\max}(N)|$ as N is symmetric. Next consider bounding the norm:

$$||N||_{\infty,\infty} = \max_{i\in S_\beta}\sum_{j\in S_\beta}|N_{ij}| \leq \frac{C_V}{2t}.$$

Obviously the last quantity becomes smaller than $\frac{C_V}{20}$, as required in Lemma 6 from Amini and Wainwright [3]. Thus we conclude that:

- $\gamma_1 = \lambda_{\max}(\widetilde{V}) \to \frac{C_V}{2}$ and, the second largest eigenvalue of $\widetilde{V}$ — γ_2, converges to o.
- The corresponding principal eigenvector of $\widetilde{V}$ — $\widetilde{z}_{S_\beta}$ satisfies the following inequality:

$$\left\|\widetilde{z}_{S_\beta} - \beta_{S_\beta}\right\|_\infty \leq \frac{1}{2\sqrt{s}}.$$

Next we show that the rest of the sign matrix U, i.e. $U_{S_\beta^c,S_\beta}$ and $U_{S_\beta^c,S_\beta^c}$ can be selected in such a way, so that the blocks $\widetilde{V}_{S_\beta^c,S_\beta}$ and $\widetilde{V}_{S_\beta^c,S_\beta^c}$ are 0. For this purpose we select $U_{S_\beta^c,S_\beta} = \frac{1}{\lambda_n}V_{S_\beta^c,S_\beta}$ and $U_{S_\beta^c,S_\beta^c} = \frac{1}{\lambda_n}V_{S_\beta^c,S_\beta^c}$. Since it is clear that the vector $(\widetilde{z}_{S_\beta}, 0_{S_\beta^c})$ is the principle eigenvector of $\widetilde{V}$, if

U is a sign matrix, Lemma 3.5.1 will conclude that — $(\widetilde{z}_{S_\beta}^\intercal, 0_{S_\beta^c}^\intercal)^\intercal(\widetilde{z}_{S_\beta}^\intercal, 0_{S_\beta^c}^\intercal)$ is the optimal solution to the optimization problem, which will in turn conclude our claim.

It remains to show that the specified U is indeed a sign matrix. Note that by Cauchy-Schwartz for $k \in S_\beta^c$ and any j, we have:

$$V^{jk} \leq \sqrt{V^{jj}}\sqrt{V^{kk}}. \tag{3.5.9}$$

From (3.5.8) if $j \in S_\beta$, we have that high probability: $V^{jj} \leq \frac{C_V}{s} + \frac{C_V}{2st} \leq \frac{3C_V}{2s}$.

Hence, it is sufficient to select m, H large enough so that: $V^{kk} \leq \frac{C_V}{6s}$, for all $k \in S_\beta^c$. Going back to (3.4.14) it can be easily seen that by selecting $x = \frac{nC_V}{36s}$, we can ensure (after applying (3.5.7)) that $V^{kk} \leq \frac{C_V}{6s}$ for all $k \in S_\beta^c$, by requiring:

$$\frac{nC_V}{36s} \geq 2\log(p-s), \tag{3.5.10}$$

from the union bound. This combined with (3.5.9) shows that the so defined matrix U is indeed a sign matrix, which concludes the proof.

□

3.6 Towards a Robust Support Recovery with Correlated Gaussian Design

In this section we consider several non-SIR based algorithms for dealing with support recovery in single index models. We will partially address the question of support recovery with generic covariance by studying the linear regression LASSO's performance.

3.6.1 Covariance Screening under $\Sigma = \mathbb{E}[XX^\intercal] = \mathbb{I}$

In this section we present another idea for signed support recovery. Let us observe n iid observations from a single index model $Y_i = f(X_i^\intercal \beta, \varepsilon_i)$, where $\varepsilon_i \perp\!\!\!\perp X_i$ and $\mathbb{E}[X_i] = 0, \mathbb{E}[X_i X_i^\intercal] = \mathbb{I}$. To this end we recall the linearity of expectation definition (which we briefly mentioned in the introduction), used in [47,46].

Definition 3.6.1. *A p-dimensional random variable X is said to satisfy linearity of expectation in the direction β if for any direction $b \in \mathbb{R}^p$:*

$$\mathbb{E}[X^\intercal b | X^\intercal \beta] = c_b X^\intercal \beta + a_b,$$

where $a_b, c_b \in \mathbb{R}$ are some real constants which might depend on the direction b.

Remark 3.6.2. *Note that if additionally $\mathbb{E}[X] = 0$, then by taking expectation it is clear that $a_b \equiv 0$.*

Evidently, linearity of expectation is direction specific by definition. However, elliptical distributions [21], are known to satisfy the linearity in expectation uniformly in all directions. We recall that a p-dimensional random variable is elliptically distributed iff its characteristic function can be written in the form $e^{it^\intercal \mu} \Psi(t^\intercal \Sigma t)$ for all $t \in \mathbb{R}^p$, for some $\mu \in \mathbb{R}^p$ and a positive definite symmetric $\Sigma \in \mathbb{R}^{p \times p}$. The function Ψ is referred to as the characteristic generator of the elliptical distribution.

One advantage of the method that we layout in this section over the SDP approach is that it doesn't require complicated optimization procedures. It requires however slightly different set of assumptions, and hence can be considered as a complement to the theory we have developed throughout this chapter. Consider the average:

$$\frac{1}{n} \sum_{i=1}^{n} Y_i X_i,$$

which is simply a vector estimating the covariance between Y and the vector X. We will study a screening procedure based on the above average, taking the s biggest absolute values of the coordinates with their corresponding signs in terms of signed support recovery.

The motivation for considering this average is by Theorem 2.1 of[47]. An application of this theorem, gives us that if X satisfies the linearity in expectation, then minimization of the problem (provided that a minimizer exists):

$$\operatorname*{argmin}_b \mathbb{E}(Y - b^\intercal X)^2 \equiv c_0\beta, \text{ for some } c_0 \in \mathbb{R}. \tag{3.6.1}$$

Under the assumption $\mathbb{E}[XX^\intercal] = \mathbb{I}$, this population version problem clearly has a unique solution of the form $[\mathbb{E}XX^\intercal]^{-1}\mathbb{E}YX = \mathbb{E}YX$ and hence we conclude that this vector is proportional to the true β. To be self-contained we include a standalone proof of this simple but important observation.

Lemma 3.6.3. *Let $X \in \mathbb{R}^p$ be a mean zero random vector, which satisfies the linearity in expectation for a direction β such that $\mathbb{E}[(X^\intercal\beta)^2] > 0$. Assume also that $\Sigma = \mathbb{E}[XX^\intercal] = \mathbb{I}$, and let $Y = f(X^\intercal\beta, \varepsilon)$ for some f and $\varepsilon \perp\!\!\!\perp X$. Then we have $\mathbb{E}[YX] = c_0\beta$, where $c_0 := \frac{\mathbb{E}[f(Z,\varepsilon)Z]}{||\beta||_2^2}$ for a random variable $Z \sim X^\intercal\beta, Z \perp\!\!\!\perp \varepsilon$.*

Proof of Lemma 3.6.3. Take any $b \perp \beta$. Note that by the linearity of expectation:

$$\mathbb{E}[\beta^\intercal XX^\intercal b | X^\intercal\beta] = c_b(X^\intercal\beta)^2.$$

Taking another expectation above, we conclude that $\mathbb{E}[\beta^\intercal XX^\intercal b] = c_b\mathbb{E}[(X^\intercal\beta)^2]$. However

$$\mathbb{E}[\beta^\intercal XX^\intercal b] = \beta^\intercal b = 0,$$

and hence $c_b = 0$. Thus we showed that if $b \perp \beta$ we have $\mathbb{E}[X^\intercal b | X^\intercal \beta] = 0$. Next, for any $b \perp \beta$ we have:

$$\mathbb{E}[Y X^\intercal b] = \mathbb{E}[\mathbb{E}[Y X^\intercal b | X^\intercal \beta]] = \mathbb{E}[\mathbb{E}[Y | X^\intercal \beta]\mathbb{E}[X^\intercal b | X^\intercal \beta]] = 0.$$

Hence $\mathbb{E}[YX] \propto \beta$. Finally, a projection on β yields the final conclusion:

$$c_0 ||\beta||_2^2 = \mathbb{E}[Y X^\intercal \beta] = \mathbb{E}[f(Z, \varepsilon) Z],$$

where $Z = X^\intercal \beta$. □

Remark 3.6.4. *The statement of Lemma 3.6.3 is readily generalizable to the situation where* $Y = f(X^\intercal \beta, \varepsilon)$*, but* $\mathbb{E}[XX^\intercal] = \Sigma > 0$ *with* $\Sigma \neq \mathbb{I}$*. This is equivalent to* $Y = f(X^\intercal \Sigma^{-1/2} \Sigma^{1/2} \beta, \varepsilon)$*, and we conclude that* $\Sigma^{-1/2}\mathbb{E}[YX] = c_0 \Sigma^{1/2} \beta$*, where* $c_0 := \frac{\mathbb{E}[f(Z,\varepsilon)Z]}{\beta^\intercal \Sigma \beta}$*, where* $Z \sim X^\intercal \beta$*.*

From Lemma 3.6.3 it follows that under the assumption $\mathbb{E} f(Z, \varepsilon) Z \neq 0$, where $Z \sim X^\intercal \beta$, the proportionality constant $c_0 \neq 0$. For identifiability we will assume that $||\beta||_2 = 1$, so that c_0 is simply $\mathbb{E} f(Z, \varepsilon) Z \neq 0$.

While in general, the assumption $\mathbb{E} f(Z, \varepsilon) Z \neq 0$[‡] is dependent on the particular direction β, note that in the case when X comes from a spherical distribution[57] (such as the normal) the projections $X^\intercal u$ have precisely the same distribution, where u is any unit vector. Recall that a random variable is spherically symmetric if it is elliptically distributed with $\Sigma = \mathbb{I}$. As we mentioned earlier, spherical distributions automatically satisfy the linearity property. We will assume henceforth that X has a spherical distribution.

In addition to the requirement of X having a spherically symmetric distribution, we will require that the coordinate-wise X has sub-Gaussian distributions. This requirement is equivalent to re-

[‡]Here, and throughout, observe that if this condition does not hold for the original function $f(Z, \varepsilon)$ it might hold for some transformation of that function. If g is such a transform, our methods will work for $g(Y)$ instead of Y.

quiring that the characteristic generator of the spherical distribution $\Psi(t) \leq \exp(-Ct)$ for all $t \in \mathbb{R}^+$, for some $C > 0$. Since we are assuming X has mean 0, by property 4 in Lemma 5.5 of[84] under the assumption that $\Psi(t) \leq \exp(-Ct), t \in \mathbb{R}^+$ we can easily conclude that coordinate-wise the distributions are sub-Gaussian in the case when $\Sigma = \mathbb{I}$. Such assumptions clearly include the case when $X \sim N(0, \mathbb{I})$, but are more generic.

Finally under regularity conditions on f and ε we will assume that the variable $f(Z, \varepsilon)$ is sub-Gaussian, where $Z \sim X^\intercal u$ with u being any unit vector. This assumption differs from the ones we assumed when analyzing the SIR approaches, and can be viewed as a disadvantage of this framework. Nevertheless, sub-Gaussianity encompasses many relevant examples — such as the linear regression, and examples where Y has finite support such as the logistic regression, which otherwise are not covered by our previous analysis due to the requirement of Y having a continuous distribution. We are now ready to state, the following simple covariance thresholding result.

Proposition 3.6.5. *Let X be a spherically distributed p dimensional random variable with $\mathbb{E}[X] = 0, \mathrm{Var}[X] = \mathbb{I}$ and characteristic function $\Psi(t^\intercal t), t \in \mathbb{R}^p, \Psi : \mathbb{R} \mapsto \mathbb{R}$, such that $\Psi(t) \leq \exp(-Ct)$ for some $C > 0$ for all $t \in \mathbb{R}^+$. Let us observe n iid copies from a single index model $Y = f(X^\intercal \beta, \varepsilon)$, where $||\beta||_2 = 1$ and $\beta^j \in \{\frac{1}{\sqrt{s}}, -\frac{1}{\sqrt{s}}, 0\}$ for all $j \in \{1, \ldots, p\}$ and some $s \in \mathbb{N}$. Assume that the function f and random variable ε satisfy $\mathbb{E}[f(Z, \varepsilon)Z] = c_0 \neq 0$, where Z has a characteristic function $\Psi(t^2), t \in \mathbb{R}$, and the random variable $f(Z, \varepsilon)$ is sub-Gaussian. Assume also that for a fixed $\Omega > 2$*

$$n \geq \frac{2\Omega^2 K}{c_0^2 \widetilde{c}} s \log p,$$

where $\widetilde{c}$ is an absolute constant, and $K = \max_{j \in \{1,\ldots,p\}} ||Y X^j||_{\psi_1}$[§]. Then the absolute value

[§]For formal definitions of ψ_1 and ψ_2 norms please see definitions (3.1.3), (3.1.4) in Chapter 3 or (4.1.2) and (4.1.3) in Chapter 4.

correlation screening recovers the signed support[¶] *with asymptotic probability* 1.

Proof of Proposition 3.6.5. Denote by $S_\beta := \{j : \beta^j \neq 0\}$ the support of the vector β. We have $|S_\beta| = s$.

Using the fact that for any two random variables S, T, we have $||ST||_{\psi_1} \leq 2||S||_{\psi_2}||T||_{\psi_2}$ we can conclude that the vectors $Y_i X_i$ are coordinate-wise sub-exponentially distributed. Denote by $K = \max_{j\in\{1,\ldots,p\}} ||YX^j||_{\psi_1}$. An application of Proposition 5.16 of[84] and the union bound then give us that:

$$\mathbb{P}\left(\left|\left|\frac{1}{n}\sum_{i=1}^{n} Y_i X_i - \mathbb{E}[YX]\right|\right|_\infty \geq t\right) \leq 2p\exp\left[-\widetilde{c}\min\left(\frac{nt^2}{K^2}, \frac{nt}{K}\right)\right],$$

where $\widetilde{c} > 0$ is some absolute constant. This inequality then gives us that

$$\sup_{j\in\{1,\ldots,p\}} \left|\frac{1}{n}\sum_{i=1}^{n} Y_i X_i^j - \mathbb{E}[YX^j]\right| \leq \frac{\sqrt{2}K}{\sqrt{\widetilde{c}}}\sqrt{\frac{\log p}{n}},$$

with probability at least $1 - 2p^{-1}$ for values of n, p such that $\frac{\log p}{n}$ is small enough. Note that, by Lemma 3.6.3 this inequality implies that if:

$$\frac{|c_0|}{\sqrt{s}} > \Omega\frac{\sqrt{2}K}{\sqrt{\widetilde{c}}}\sqrt{\frac{\log p}{n}} \text{ for any } \Omega > 2,$$

there will be a gap in the absolute values of the coefficients of $|\frac{1}{n}\sum_{i=1}^{n} Y_i X_i^j|$ for $j \in S_\beta$ and $j \notin S_\beta$. This is because:

$$\frac{|c_0|}{\sqrt{s}} - \frac{\sqrt{2}K}{\sqrt{\widetilde{c}}}\sqrt{\frac{\log p}{n}} \geq (\Omega - 1)K\frac{\sqrt{2}}{\sqrt{\widetilde{c}}}\sqrt{\frac{\log p}{n}} > \frac{\sqrt{2}K}{\sqrt{\widetilde{c}}}\sqrt{\frac{\log p}{n}}.$$

This also shows that the coefficients will achieve the correct sign. We conclude that as long as $n \geq$

[¶] By signed support we mean recovering the signed support of $\text{sign}(c_0\beta)$.

$\frac{2\Omega^2 K}{c_0^2 \widetilde{c}} s \log p$, sign detection is possible. □

Relaxing the sub-Gaussianity of $Y = f(Z, \varepsilon)$

In what follows we look into relaxing the requirement on sub-Gaussianity on the Y distribution, to allow for more heavy tailed distributions of the outcome. In order to do so, we will impose a more stringent restriction on the X distribution, namely we will assume that $X \sim N(0, \mathbb{I})$. Next we proceed to show:

Proposition 3.6.6. *Let $X \sim N(0, \mathbb{I})$ be a p dimensional random variable. Let us observe n iid copies from a single index model $Y = f(X^\intercal \beta, \varepsilon)$, where $||\beta||_2 = 1$ and $\beta^j \in \{\frac{1}{\sqrt{s}}, -\frac{1}{\sqrt{s}}, 0\}$ for all $j \in \{1, \ldots, p\}$ and some $s \in \mathbb{N}$. Consider the a function f and random variable ε such that $\mathbb{E}[f(Z, \varepsilon)Z] = c_0 \neq 0$, where $Z \sim N(0, 1)$, and let $\sigma^2 := \mathbb{E}(f(Z, \varepsilon)^2) < \infty$, $\eta := \mathrm{Var}(f^2(Z, \varepsilon)) < \infty$ and $\gamma := \mathrm{Var}[f(Z, \varepsilon)Z] < \infty$. Then as long as $n \geq \frac{81}{c_0^2}(\sigma^2 + 1) s \log p$ the absolute value correlation screening recovers the signed support with asymptotic probability* 1.

Proof of Proposition 3.6.6. We follow the same steps as the proof of Proposition 3.6.5. We will use the following Lemma which we show in the appendix:

Lemma 3.6.7. *Let us observe n data points from the model described in Proposition 3.6.6 with β being an arbitrary unit vector. Then we have that with probability at least $1 - \frac{\eta+\gamma}{\log n} - \frac{2}{p}$ the following event holds:*

$$\left|\left| \frac{1}{n} \sum_{i=1}^{n} Y_i X_i - E[YX] \right|\right|_\infty \leq \frac{||\beta||_\infty \sqrt{\log n}}{\sqrt{n}} + 2\sqrt{(\sigma^2 + 1)\frac{\log p}{n}}.$$

Using the fact that $E[YX] = c_0 \beta$ in our case and that $||\beta||_\infty = \frac{1}{\sqrt{s}}$, we have that if:

$$|c_0| \sqrt{n} \geq 4\sqrt{\log n} + 8\sqrt{\sigma^2 + 1}\sqrt{s \log p},$$

there will have a gap between the coefficients. Note that this condition holds if $16 \log n \leq (\sigma^2 + 1)s \log p$, by our assumption. If this doesn't hold then the inequality above holds trivially for large enough n. □

3.6.2 A Solution Towards a Generic Covariance Structure

In this section we will consider the more general problem, where we observe n samples $Y_i = f(X_i^\intercal \beta^*, \varepsilon_i)$, $i = 1, \ldots, n$, but the distribution $X_i \sim N(0, \Sigma)$ where the covariance matrix Σ is unknown. We will use matrix notation. For convenience we denote with bold script the $n \times p$ matrix $\mathbf{X}$ whose rows are the vectors $X_i^\intercal, i = 1, \ldots, n$. By indexing the matrix with a set $A \subset \{1, \ldots, p\}$ (including a single index) $\mathbf{X}_A$ we mean taking only predictors corresponding to the set A and concatenating them. We denote with $\mathbf{Y}$ the concatenation of values $Y_i, i = 1, \ldots, n$.

Under certain sufficient conditions, our goal is to show that the LASSO algorithm recovers the support of the vector β with asymptotic probability 1. For identifiability we will work under the scenario $\beta^{*\intercal} \Sigma \beta^* = 1$. In this section we prefer working slightly more generally and we will not require each of the signals in β to be of the same magnitude. Inspired by Section 3.6.1 we define the observed residual:

$$\mathbf{w} = \mathbf{Y} - c_0 \mathbf{X} \beta^*,$$

where just as before we have:

$$c_0 := \mathbb{E}[Y X^\intercal \beta^*] = \mathbb{E}[f(Z, \varepsilon) Z], \text{ for } Z \sim N(0, 1) \tag{3.6.2}$$

Note that $\mathbf{w}$ is not mean 0, but on the other hand by Remark 3.6.4 we have $\mathbb{E}[\mathbf{X}^\intercal \mathbf{w}] = 0$. In terms of the $\mathbf{w}$ notation, we can also write trivially $\mathbf{Y} = c_0 \mathbf{X} \beta^* + \mathbf{w}$. In this section we are interested in studying the support recovery properties of a vector obtained through solving the following pro-

gram:

$$\widehat{\beta} = \underset{\beta \in \mathbb{R}^p}{\operatorname{argmin}} \frac{1}{2n} ||\mathbf{Y} - \mathbf{X}\beta||_2^2 + \lambda ||\beta||_1, \tag{3.6.3}$$

also known as Linear Regression LASSO.

The most detailed results regarding the support consistency of the LASSO in the linear model up to date, to the best of our knowledge can be found in [87]. Seminal analysis of the problem was performed by [62,98]. We would like to stress the fact that we are dealing with a more general problem, with the data being generated through a single index model rather than the usual linear model.

Next we summarize a primal dual witness (PDW) construction which we borrow from [87]. The PDW construction lays out steps allowing one to prove sign consistency for L_1 constrained quadratic programming (3.6.3). We will only provide the sufficient conditions to show sign-consistency, and the interested reader can check [87] for the necessary conditions. We note that the proof of the PDW construction is generic, in that it does not rely on the distribution of $\mathbf{w}$, and hence extends to the current framework.

For a vector $v \in \mathbb{R}^p$ let $S(v) = \{i : v_i \neq 0\}$, and let $S = S(\beta^*)$ for brevity. As we mentioned we are interested, more generally, in signed support recovery. Define the signed support $S_{\pm}(v) = \{\text{sign}(v_i)\}_{i=1}^p$ where $\text{sign}(0) = 0$.

Recall that a vector z is a subgradient of the L_1 norm evaluated at a vector $v \in \mathbb{R}^p$ (i.e. $z \in \partial||v||_1$) if we have $z_i = \text{sign}(v_i), v_i \neq 0$ and $z_i \in [-1, 1]$ otherwise. It follows from Karush-Kuhn-Tucker's theorem that a vector $\widehat{\beta} \in \mathbb{R}^p$ is optimal for the LASSO problem (3.6.3) iff there exists a subgradient $\widehat{z} \in \partial||\widehat{\beta}||_1$ such that:

$$\frac{1}{n}\mathbf{X}^\intercal\mathbf{X}(\widehat{\beta} - c_0\beta^*) - \frac{1}{n}\mathbf{X}^\intercal\mathbf{w} + \lambda\widehat{z} = 0. \tag{3.6.4}$$

We will assume that the matrix $\mathbf{X}_S^\intercal \mathbf{X}_S$ is invertible, even though this is not required by the PDW. The PDW method constructs a pair $(\check{\beta}, \check{z}) \in \mathbb{R}^p \times \mathbb{R}^p$ by following the steps:

- Solve:

$$\check{\beta}_S = \arg\min_{\beta_S \in \mathbb{R}^s} \frac{1}{2n} ||\mathbf{Y} - \mathbf{X}_S \beta_S||_2^2 + \lambda ||\beta_S||_1,$$

  where $s = |S|$. This solution is unique under the invertibility of $\mathbf{X}_S^\intercal \mathbf{X}_S$. Set $\check{\beta}_{S^c} = 0$.

- Choose $\check{z}_S$ to be in $\partial ||\check{\beta}_S||_1$.

- For $j \in S^c$ set $Z_j := \mathbf{X}_j^\intercal [\mathbf{X}_S (\mathbf{X}_S^\intercal \mathbf{X}_S)^{-1} \check{z}_S + P_{\mathbf{X}_S^\perp} \left(\frac{\mathbf{w}}{\lambda n}\right)]$[‖], where $P_{\mathbf{X}_S^\perp} = \mathbb{I} - \mathbf{X}_S (\mathbf{X}_S^\intercal \mathbf{X}_S)^{-1} \mathbf{X}_S^\intercal$ is an orthogonal projection. Checking that $|Z_j| < 1$ for all $j \in S^c$ ensures that there is a unique solution $\check{\beta} = (\check{\beta}_S^\intercal, \check{\beta}_{S^c}^\intercal)^\intercal$ satisfying $S(\check{\beta}) \subseteq S(c_0 \beta^*)$. Verifying that $|Z_j| < 1$ is referred to as verifying *strict dual feasibility*.

- To check sign consistency we need $\check{z}_S = \text{sign}(c_0 \beta_S^*)$. For each $j \in S$, define:

$$\Delta_j := e_j^\intercal (n^{-1} \mathbf{X}_S^\intercal \mathbf{X}_S)^{-1} \left[n^{-1} \mathbf{X}_S^\intercal \mathbf{w} - \lambda \, \text{sign}(c_0 \beta_S^*) \right],$$[**]

  where $e_j \in \mathbb{R}^s$ is a unit vector with 1 at the j^{th} position. Checking $\check{z}_S = \text{sign}(c_0 \beta_S^*)$ is equivalent to checking:

$$\text{sign}(c_0 \beta_i^* + \Delta_i) = \text{sign}(c_0 \beta_i^*), \; \forall i \in S.$$

To this end we require several restrictions on the covariance matrix. We partition the covariance matrix $\Sigma = \begin{bmatrix} \Sigma_{SS} & \Sigma_{SS^c} \\ \Sigma_{S^c S} & \Sigma_{S^c S^c} \end{bmatrix}$, where Σ_{SS} corresponds to the covariance of X_S.

[‖] Z_j are derived by simply plugging in $\check{\beta}$ and $\check{z}_S$ and solving (3.6.4) for $\check{z}_{S^c}$.

[**] Δ_j can be seen to be equal to $\check{\beta}_j - c_0 \beta_j^*$ for $j \in S$, when $\check{z}_S = \text{sign}(c_0 \beta^*)$.

Assumption 3.6.8 (Irrepresentable Condition). *Assume that:*

$$||\Sigma_{S^cS}\Sigma_{SS}^{-1}||_{\infty,\infty} \leq (1-\kappa),$$

for some $\kappa > 0$.

Assumption 3.6.9 (Bounded Spectrum). *Let*

$$\lambda_{\min}^S \leq \Sigma_{SS} \leq \lambda_{\max}^S,$$

for some fixed $0 < \lambda_{\min}^S \leq \lambda_{\max}^S < \infty$.

Next, we define several shorthand notations which we will use in our main result. Let:

$$\Sigma_{S^c|S} := \Sigma_{S^cS^c} - \Sigma_{S^cS}\Sigma_{SS}^{-1}\Sigma_{SS^c}, \tag{3.6.5}$$

$$\rho_\infty(\Sigma_{SS}^{1/2}) := ||\Sigma_{SS}^{-1/2}||_{\infty,\infty}||\Sigma_{SS}^{1/2}||_{\infty,\infty}, \tag{3.6.6}$$

be the conditional covariance matrix of $X_{S^c}|X_S$, and the condition number of $\Sigma_{SS}^{1/2}$ with respect to $||\cdot||_{\infty,\infty}$ correspondingly.

Furthermore we will need the following quantities:

$$\sigma^2 := \mathbb{E}[f(Z,\varepsilon)^2],\ \eta := \mathrm{Var}[f^2(Z,\varepsilon)],\ \gamma := \mathrm{Var}[f(Z,\varepsilon)Z] \tag{3.6.7}$$

$$\xi^2 := \mathbb{E}[(f(Z,\varepsilon) - c_0 Z)^2],\ \theta^2 := \mathrm{Var}[(f(Z,\varepsilon) - c_0 Z)^2] \tag{3.6.8}$$

where $Z \sim N(0,1)$. In order for all these moments to be well defined we need the following:

Assumption 3.6.10 (Bounded 4^{th} Moment)**.** *We assume that:*

$$\mathbb{E}[f(Z,\varepsilon)^4] < \infty. \tag{3.6.9}$$

Assumption 3.6.10 guarantees that all of the shorthand notations defined in (3.6.7) and (3.6.8) are well defined and finite. Finally, successful support recovery will depend on the strength of the minimal signal in β^*. Let $||\beta^*||_{\min} := \min_{i\in S} |\beta_i^*|$, be the minimal non-zero signal in the vector β^*

We are now ready to provide sufficient conditions for the LASSO signed support recovery, in the setting of single index models:

Theorem 3.6.11. *Assume that Assumptions 3.6.8–3.6.10 hold. Let $Y_i = f(\beta^{*\intercal} X_i, \varepsilon_i)$ with $X_i \sim N(0,\Sigma), i = 1,\dots,n$, be iid data generated from a single index model. Let $\widehat{\beta}$ be the solution to the optimization program defined in (3.6.3), with λ being a tuning parameter. Assume furthermore $s = O(p^{1-\omega})$ for some $\omega > 0$. Then we have the following sufficient conditions:*

i. If

$$n \geq \frac{4\log(p-s) d_{\max}(\Sigma_{S^c|S}) \left(\frac{4s}{\lambda^S_{\min}} + \frac{\xi^2+1}{\lambda^2}\right)}{\kappa^2},$$

then $S(\widehat{\beta}) \subseteq S(c_0\beta^*)$, with probability at least $1 - \frac{2}{p-s} - \frac{\theta^2}{n} - 2\exp(-s/2)$.

ii. There exist some absolute constants $\Omega_0, \Omega_1, \Omega_2, \Omega_3 > 0$ which may depend on c_0, σ such that if:

$$||\beta^*||_{\min} \geq ||\Sigma_{SS}^{-1/2}||^2_{\infty,\infty}\lambda\Omega_0 + \rho_\infty(\Sigma_{SS}^{1/2})||\beta^*||_\infty \left[\Omega_1\sqrt{\frac{s\log(p-s)}{n}} + \Omega_2\sqrt{\frac{\log n}{n}}\right]$$
$$+ ||\Sigma_{SS}^{-1/2}||_{\infty,\infty}\Omega_3\sqrt{\frac{\log s}{n}},$$

we have $S_{\pm}(\widehat{\beta}) = S_{\pm}(c_0\beta^*)$ *with probability at least* $1 - 12\exp(-C_2\min(s, \log(p - s))) - 6\exp(-s/2) - \frac{6+\theta^2}{n} - \frac{2}{p-s} - \frac{8}{s} - 2\frac{\eta+\gamma}{\log n}$, *where* $C_2 > 0$ *is an absolute constant.*

Before we proceed with the proof of our statement we would like to mention a few remarks on our sufficient conditions, in particular the ones suggested in ii.

Remark 3.6.12. *Observe that* $\frac{1}{\lambda^S_{\max}} \leq ||\beta^*||_2 \leq \frac{1}{\lambda^S_{\min}}$. *Hence the value of* $||\beta^*||_{\min}$ *is of asymptotically "largest" order when* $||\beta^*||_{\min} \asymp ||\beta^*||_\infty \asymp \frac{1}{\sqrt{s}}$. *Setting*

$$\lambda := \lambda_T = \sqrt{(\xi^2 + 1)\frac{4C_T d_{\max}(\Sigma_{S^c|S})}{\kappa^2}\frac{\log(p-s)}{n}},$$

for some $C_T > 1$ *gives us that the condition from i. is equivalent to:*

$$\frac{n}{s\log(p-s)} \geq \frac{16 d_{\max}(\Sigma_{S^c|S})}{(1 - C_T^{-1})\kappa^2 \lambda^S_{\min}}.$$

Note that due to positive definiteness: $d_{\max}(\Sigma_{S^c|S}) \leq \lambda^S_{\max}$, *and hence is a bounded quantity by assumption. Assume additionally* $||\Sigma_{SS}^{-1/2}||^2_{\infty,\infty} = O(1)$, $\rho_\infty(\Sigma_{SS}^{1/2}) = O(1)$, *and let* $||\beta^*||_{\min} \asymp \frac{1}{\sqrt{s}}$. *Using the same* $\lambda = \lambda_T$ *we can clearly achieve the sufficient condition in ii. by potentially over-scaling the ratio between* $\frac{n}{s\log(p-s)}$. *On the other hand, this scaling can no longer be guaranteed if* $||\beta^*||_{\min} \asymp \frac{1}{\sqrt{s}}$ *fails to hold.*

Proof of Theorem 3.6.11. Our proof follows Theorem 3 in[87]. For completeness and to increase readability, we will provide a full proof of this theorem, while explicitly stating where modifications of the original argument were required.

Verifying Strict Dual Feasibility

For $j \in S^c$ decompose $\mathbf{X}_j^\intercal = \Sigma_{jS}\Sigma_{SS}^{-1}\mathbf{X}_S^\intercal + E_j^\intercal$, where the elements of the prediction error vector $E_j \in \mathbb{R}^n$ are iid with $E_{ij} \sim N(0, [\Sigma_{S^c|S}]_{jj}), i = 1, \ldots, n$. Following the definition of Z_j gives us

that $Z_j = A_j + B_j$, where:

$$A_j := E_j^\intercal \left[\mathbf{X}_S (\mathbf{X}_S^\intercal \mathbf{X}_S)^{-1} \check{z}_S + P_{\mathbf{X}_S^\perp} \left(\frac{\mathbf{w}}{\lambda n} \right) \right], \tag{3.6.10}$$

$$B_j := \Sigma_{jS} (\Sigma_{SS})^{-1} \check{z}_S. \tag{3.6.11}$$

Under the irrepresentable condition, we have that $\max_{j \in S^c} |B_j| \leq (1 - \kappa)$. Conditional on $\mathbf{X}_S$ and $\boldsymbol{\varepsilon}$ (which determine $\mathbf{w} = \mathbf{Y} - c_0 \mathbf{X} \beta^*$) we have that the gradient $\check{z}_S$ is independent of the vector E_j because the gradient is deterministic after conditioning on these quantities [††]. We have that $\mathrm{Var}(E_{ij}) \leq d_{\max}(\Sigma_{S^c|S})$, and thus conditionally on $\mathbf{X}_S$ and $\boldsymbol{\varepsilon}$ we get:

$$\begin{aligned} \mathrm{Var}(A_j) &\leq d_{\max}(\Sigma_{S^c|S}) \left|\left| \mathbf{X}_S (\mathbf{X}_S^\intercal \mathbf{X}_S)^{-1} \check{z}_S + P_{\mathbf{X}_S^\perp} \left(\frac{\mathbf{w}}{\lambda n} \right) \right|\right|_2^2 \\ &= d_{\max}(\Sigma_{S^c|S}) \left[\check{z}_S^\intercal (\mathbf{X}_S^\intercal \mathbf{X}_S)^{-1} \check{z}_S + \left|\left| P_{\mathbf{X}_S^\perp} \left(\frac{\mathbf{w}}{\lambda n} \right) \right|\right|_2^2 \right]. \end{aligned}$$

Next we formulate a lemma which is a slight modification of Lemma 4 in [87]. The reason for this modification is that in our case $\mathbf{w}$ is no longer $\sim N(0, \sigma^2 \mathbb{I})$.

Lemma 3.6.13. *Assume that* $\frac{s}{n} \leq \frac{1}{16}$. *Then we have:*

$$\max_{j \in S^c} \mathrm{Var}(A_j) \leq \underbrace{d_{\max}(\Sigma_{S^c|S}) \left(\frac{4s}{\lambda_{\min}^S n} + \frac{\xi^2 + 1}{\lambda^2 n} \right)}_{M},$$

with probability at least $1 - 2\exp(-s/2) - \frac{\theta^2}{n}$.

Now since conditionally on $\mathbf{X}_S$ and $\boldsymbol{\varepsilon}$ we have $A_j \sim N(0, \mathrm{Var}(A_j))$, using a standard normal

[††] Observe that by (3.6.4), we have $\check{z}_S = -\frac{1}{n} \boldsymbol{X}_S^\intercal \boldsymbol{X}_S (\check{\beta}_S - c_0 \beta^*) + \frac{\boldsymbol{X}_S^\intercal \mathbf{w}}{\lambda n}$

tail bound and the union bound we conclude:

$$\mathbb{P}(\max_{j\in S^c}|A_j|\geq\kappa)\leq 2(p-s)\exp(-\kappa^2/(2M))+2\exp(-s/2)+\frac{\theta^2}{n}.$$

We need to select M so that the exponential term is decaying in the above display. A sufficient condition for this is $\kappa^2/(2M)\geq 2\log(p-s)$. The last is equivalent to:

$$n\geq\frac{4\log(p-s)d_{\max}(\Sigma_{S^c|S})\left(\frac{4s}{\lambda^S_{\min}}+\frac{\xi^2+1}{\lambda^2}\right)}{\kappa^2}.$$

Verifying Sign Consistency

The last part shows that the LASSO has a unique solution $\widehat{\beta}$ which satisfies $S(\widehat{\beta})\subseteq S(c_0\beta^*)$ with high probability. Now we need to verify the sign-consistency, in order to show that the supports will coincide. We have the following:

$$\max_{i\in S}|\Delta_i|\leq\underbrace{\lambda\left|\left|(n^{-1}\mathbf{X}_S^\intercal\mathbf{X}_S)^{-1}\operatorname{sign}(c_0\beta^*_S)\right|\right|_\infty}_{I_1}+\underbrace{\left|\left|(\mathbf{X}_S^\intercal\mathbf{X}_S)^{-1}\mathbf{X}_S^\intercal\mathbf{w}\right|\right|_\infty}_{I_2}.$$

To deal with the first term we need the following:

Lemma 3.6.14. *There exist positive constants $K_1, C_2>0$, such that the following holds:*

$$\mathbb{P}(I_1\geq\lambda K_1||\Sigma_{SS}^{-1/2}||^2_{\infty,\infty})\leq 4\exp(-C_2\min(s,\log(p-s))),$$

The proof of this lemma is part of the proof of Theorem 3 in [87]. Next we turn to bounding the term I_2. This is where our proof departs substantially from the proof in [87], as I_2 no longer has a simple structure required in the original argument. In our case $\mathbf{w}$ depends on $\mathbf{X}_S$, and it is not mean 0. We will make usage of the following result, whose proof is provided in the appendix:

Lemma 3.6.15. *Let* $||\beta^*||_2 = 1$. *We have* n *iid observations* $Y = f(\beta^{*\intercal}X, \varepsilon)$ *from a single index model, where* $X \sim N(0, \mathbb{I}_{s\times s})$, *with* $s < n$. *Then there exist some absolute constants* $\Omega_1, \Omega_2, \Omega_3 > 0$ *(depending on* σ *and* $|c_0|$*), such that:*

$$||[\mathbf{X}^\intercal\mathbf{X}]^{-1}\mathbf{X}^\intercal\mathbf{Y} - c_0\beta^*||_\infty \leq \Omega_1||\beta^*||_\infty\sqrt{\frac{s\log(p-s)}{n}} + \Omega_2||\beta^*||_\infty\sqrt{\frac{\log n}{n}} + \Omega_3\sqrt{\frac{\log s}{n}},$$

with probability at least $1 - 8\exp(-C_2\min(s, \log(p-s))) - 4\exp(-s/2) - \frac{6}{n} - \frac{8}{s} - 2\frac{\eta+\gamma}{\log n}$, *where* $C_2 > 0$ *is the same absolute constant as in Lemma 3.6.14. Denote for brevity the RHS of the inequality as* $\delta(||\beta||_\infty, n, s, p)$.

While Lemma 3.6.15 is stated in terms of standard multivariate normal distribution $N(0, \mathbb{I})$, we can easily adapt it to more general situations where we observe non-standard normal random variables $N(0, \Sigma_{SS})$. Next recall that the rows of $\mathbf{X}_S$ are distributed as $N(0, \Sigma_{SS})$, $Y_i = f(\beta^\intercal X_i, \varepsilon)$, and $\beta_S^\intercal\Sigma_{SS}\beta_S = 1$. Denote with $\mathbf{Z} = \Sigma_{SS}^{-1/2}\mathbf{X}_S^\intercal$. Then we have the following inequality, with high probability:

$$\begin{aligned} I_2 = ||[\mathbf{X}_S^\intercal\mathbf{X}_S]^{-1}\mathbf{X}_S^\intercal\mathbf{Y} - c_0\beta_S^*||_\infty &= ||\Sigma_{SS}^{-1/2}[\mathbf{Z}^\intercal\mathbf{Z}]^{-1}\mathbf{Z}^\intercal\mathbf{Y} - c_0\beta_S^*||_\infty \\ &\leq ||\Sigma_{SS}^{-1/2}||_{\infty,\infty}||[\mathbf{Z}^\intercal\mathbf{Z}]^{-1}\mathbf{Z}^\intercal\mathbf{Y} - c_0\Sigma_{SS}^{1/2}\beta_S^*||_\infty \\ &\leq ||\Sigma_{SS}^{-1/2}||_{\infty,\infty}\delta(||\Sigma_{SS}^{1/2}\beta_S^*||_\infty, s, n, p). \end{aligned}$$

The last two inequalities imply that:

$$\max_{i\in S}|\Delta_i| \leq \lambda K_1||\Sigma_{SS}^{-1/2}||_{\infty,\infty}^2 + ||\Sigma_{SS}^{-1/2}||_{\infty,\infty}\delta(||\Sigma_{SS}^{1/2}\beta_S^*||_\infty, s, n, p).$$

Hence as long as for $||\beta^*||_{\min} = \min\{|\beta_i^*| : i \in S\}$ we have:

$$|c_0|||\beta||_{\min} \geq \lambda K_1 ||\Sigma_{SS}^{-1/2}||_{\infty,\infty}^2 + ||\Sigma_{SS}^{-1/2}||_{\infty,\infty} \delta(||\Sigma_{SS}^{1/2}||_{\infty,\infty} ||\beta^*||_\infty, s, n, p),$$

the LASSO will recover the support with high-probability. This concludes the proof. □

Numerical Results

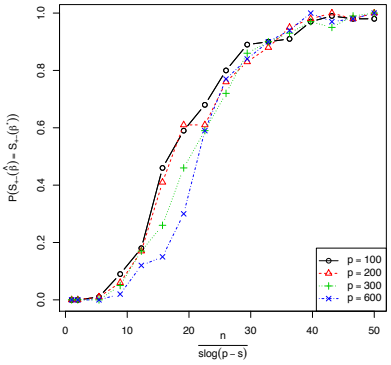
To support our theoretical claims, and in particular Theorem 3.6.11 we provide brief numeric analysis in this section. We consider the same 4 models as in the case of SIR – models (3.3.1), (3.3.2), (3.3.3) and (3.3.3).

We used a Toeplitz covariance matrix for the simulations with $\Sigma_{ij} = \frac{1}{2^{|i-j|}}$. The vector β^* was selected so that $\beta^{*\intercal}\Sigma\beta^* = 1$, the entries were equal, with the first one having a negative sign, and the rest being positive. Note that Toeplitz matrices can be seen to satisfy the requirements we impose in Section 3.6.2.

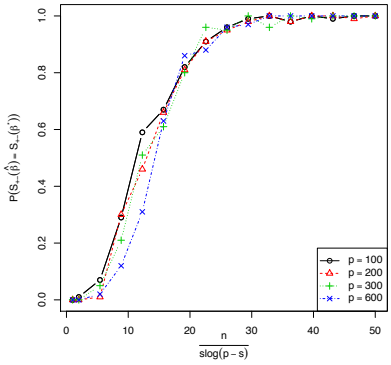
We did not tune the tuning parameter λ, but rather the selection was based on selecting the vector β on the solution path of the LASSO which contains exactly s elements. This method is justified as our theory shows the existence of a λ on the solution path recovering the correct support.

In figure 3.4, we present results of signed support recovery for different p values in the regime $s = \sqrt{p}$.

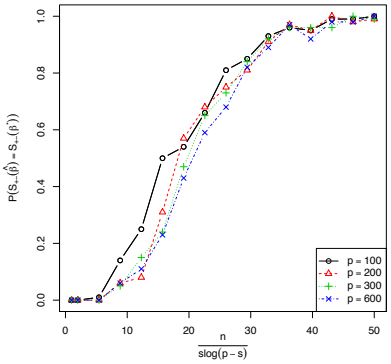
Figure 3.4: Linear Regression LASSO, $s = \sqrt{p}$



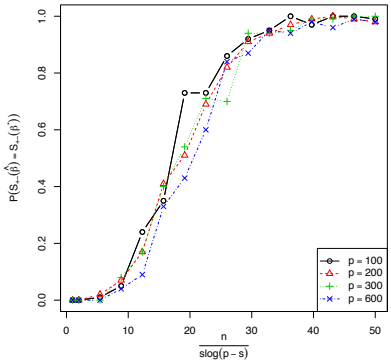
(a) Model (3.3.1)



(b) Model (3.3.2)



(c) Model (3.3.3)



(d) Model (3.3.4)

These plots illustrate different phase transitions occurring for the four different models. We observe empirically that the value of the phase transition parameter can be quite large, and hence we might be able to appreciate the effect of the scaling provided in Theorem 3.6.11 only asymptotically.

3.7 Discussion

In this chapter we studied support recovery for SIR in a high dimensional setting, under the assumption that $X \sim N(0, \mathbb{I})$. We showed that two algorithms DT and SDP, originally suggested in the sparse PCA literature, recover the support with an optimal sample size up to a multiplicative constant. To the best of our knowledge, this phenomenon has not been pointed out in the present literature. We furthermore, pointed out interestingly that the number of slices H does not need to diverge to ∞ as long as it is large enough for the DT algorithm to work.

We note that the rather restrictive assumption $X \sim N(0, \Sigma)$, can be easily extended to cover matrices Σ, which have $\Sigma_{S_\beta, S_\beta} = \mathbb{I}_{s\times s}$, $\Sigma_{S_\beta, S_\beta^c} = 0$, $\lambda_{\max}(\Sigma_{S_\beta^c, S_\beta^c}) \leq 1$. A more refined extension of the covariance structure, is not straightforward however, and warrants future research. Other extensions of this work, that we are currently working on, include more than one-dimensional SDR spaces. Our results are also motivating us to study the minimax rate for SIR under similar conditions.

In addition we considered non-SIR based approaches for support recovery, such as the covariance thresholding and the linear regression LASSO algorithm. We showed that under slightly different assumptions covariance thresholding can produce signed support recovery with a sample size of the same order as SIR. In addition, we saw that the linear regression LASSO works for variable selection with correlated Gaussian designs, even under the more general setting of single index models, provided that certain sufficient conditions (most notably the irrepresentable condition) are met.

I only believe in statistics that I doctored myself.

attributed to Winston Churchill by Joseph Goebbels

4

A Unified Theory for Inference in High-Dimensional Estimating Equations

4.1 Introduction

We are given n independent and identically distributed random samples $\{\boldsymbol{X}_i \in \mathbb{R}^q\}_{i=1,\ldots,n}$ from a statistical model $\mathcal{P} = \{\mathbb{P}_{\boldsymbol{\beta}} : \boldsymbol{\beta} \in \boldsymbol{\Omega}\}$, where $\boldsymbol{\beta} \in \mathbb{R}^d$ is an unknown parameter with $d \gg n$.

Assume that the true parameter $\boldsymbol{\beta}^*$ can be determined uniquely by solving an equation system $\mathbb{E}\mathbf{h}(\boldsymbol{X}, \boldsymbol{\beta}) = 0$, where $\mathbf{h} : \mathbb{R}^q \times \mathbb{R}^d \mapsto \mathbb{R}^d$ is a system of estimating equations and $\boldsymbol{X} \sim \mathbb{P}_{\boldsymbol{\beta}^*}$. When $d > n$, directly solving a sample version of these estimating equations is an ill-posed problem . To avoid this problem, a popular approach is to impose sparsity assumption on $\boldsymbol{\beta}^*$, which motivates large families of constrained Z-estimators in a generic form[14]:

$$\widehat{\boldsymbol{\beta}} = \text{argmin} \, \|\boldsymbol{\beta}\|_1, \text{ subject to } \left\| n^{-1} \sum_{i=1}^{n} \mathbf{h}(\boldsymbol{X}_i, \boldsymbol{\beta}) \right\|_\infty \leq \lambda, \tag{4.1.1}$$

where λ is a regularization parameter controlling the bias and variance tradeoff. Let $\boldsymbol{\beta} = (\theta, \boldsymbol{\gamma}^T)^T$, where θ is a univariate parameter of interest and $\boldsymbol{\gamma}$ is a $(d-1)$-dimensional nuisance parameter. We aim to to test the hypothesis $H_0 : \theta^* = 0$ and obtain valid confidence regions for θ^*.

As an example, consider the special case when $h((Y, \boldsymbol{Z}), \boldsymbol{\beta}) = \boldsymbol{Z}(\boldsymbol{Z}^T\boldsymbol{\beta} - Y)$, where $X = (Y, \boldsymbol{Z}^T)^T$ with $Y \in \mathbb{R}$ being the response and $\boldsymbol{Z} \in \mathbb{R}^d$ being the predictor variables. In this case formulation (4.1.1) reduces to the Danzig Selector estimator[14]. While both oracle properties Bickel et al.[9], Candes and Tao[14], Koltchinskii et al.[40] and model selection consistency results Gai et al.[25], Ye and Zhang[91], Wainwright[87] have been established for the Dantzig Selector, hypothesis testing and construction of confidence regions for the parameters have not been well explored. The main challenge in these two inferential problems, in contrast to the conventional fixed d setting, is the fact that the dimension of the nuisance parameter $\boldsymbol{\gamma}$ can be very high especially in the case $d \gg n$. In this chapter we argue that it is indeed possible to achieve the two inferential goals under some further sparsity assumptions of a certain covariance operator on the X distribution.

Such a generic framework has surprisingly many applications. For instance, consider the setting when the true parameter can be determined through minimizing a convex and sufficiently smooth loss function $\ell : \mathbb{R}^q \times \mathbb{R}^d \mapsto \mathbb{R}$, i.e. $\boldsymbol{\beta}^* = \text{argmin}_{\boldsymbol{\beta}} \mathbb{E}\ell(\boldsymbol{X}, \boldsymbol{\beta})$ with $\boldsymbol{X} \sim \mathbb{P}_{\boldsymbol{\beta}^*}$. In such cases one can equivalently solve the equation $\mathbb{E}\mathbf{h}(\boldsymbol{X}, \boldsymbol{\beta}) = 0$ where $\mathbf{h} = \frac{\partial \ell}{\partial \boldsymbol{\beta}}$, and hence infernce on

many high-dimensional M-estimators can be addressed through our framework. Moreover, there are a lot of existing constrained Z-estimators, which naturally belong in our framework. Such estimators include the Dantzig Selector[14], the CLIME estimator for inverse covariance matrices[13], sparse linear discriminant analysis (LDA) with the LDP algorithm[12] and vector autoregressive models[29]. Performing inference for the CLIME estimator has implications in graphical modeling. If the data is Gaussian, then such hypothesis testing is equivalent to edge testing in the graph structure. More generally our framework can be used to perform inference for Transelliptical graphical models, suggested by Liu et al.[52]. To the best of our knowledge, non of the aforementioned algorithms has been equipped with inferential procedures.

In order for us to construct test statistic and confidence regions, we project the estimating equation onto a certain sparse direction. We demonstrate that in doing so, one eliminates the influence of the nuisance parameter and the test statistic achieves asymptotic normality under the null hypothesis. Under more stringent conditions, we further establish uniform weak convergence to a normal distribution over a sufficiently sparse parameter set, given the null hypothesis holds. Moreover, we study the local power of our prosed test statistic and demonstrate that the same transition as in the low dimensional case occurs. In order to construct confidence regions, we suggest a corrected version $\widetilde{\theta}$ of the estimator $\widehat{\theta}$. We show that the asymptotic distribution of $\widetilde{\theta}$ is asymptotically normal, and in addition the asymptotic variance of the estimator coincides with the one used to normalize the test statistic. This demonstrates the asymptotic equivalence of the suggested confidence regions and hypothesis tests. Furthermore, in settings when $\mathbf{h} = \frac{\partial \ell}{\partial \boldsymbol{\beta}}$ with ℓ being the log-likelihood function, our estimator $\widetilde{\theta}$ achieves the optimal lower bound on the variance of over all unbiased estimates.

4.1.1 Connections with Related Work

In a parallel line of work, where a likelihood function is available in a high dimensional parametric model, one could opt for estimating the sparse parameter β through a penalized likelihood. The archetypical example of such approach in the linear model is the LASSO [78]. The theoretical properties of the LASSO have also been successfully studied in the literature, and the interested reader can look into [9,11,39], to name a few references. More generally, theoretical guarantees for solving penalized M-estimators can be found in [56]. Although estimation of β has been studied well in the high-dimensional setting, the question how to perform inference remains largely unanswered. In particular, Knight and Fu [39] showed that the asymptotic distribution of the LASSO estimator is not normal even in settings where $d < n$. There have been several different propositions how to address this question in the linear model case. P-values and confidence intervals based on sample splitting and subsampling were suggested by Meinshausen et al. [64], Meinshausen and Bühlmann [63], Shah and Samworth [72], Wasserman and Roeder [89]. For the LASSO estimator, Lockhart et al. [55], Taylor et al. [75], Lee et al. [44] suggested conditional tests based on covariates which have been selected by the LASSO. We stress the fact that this type of tests are of fundamentally different nature compared to our work. In the linear and logistic model cases [7,8] proposed an instrumental variable and double selection procedures correspondingly to produce asymptotically normal estimators. Coming from a different reasoning Zhang and Zhang [96], Javanmard and Montanari [36], van de Geer et al. [80] proposed a low dimension projection estimator, debiasing and desparsifying correction methods correspondingly, for constructing confidence intervals in high dimensional models with the L_1 penalty. Recently in a related framework, Ning and Liu [69] proposed a projected score test in a semi parametric high dimensional setting, which works for a wider class of penalty functions. A different score related approach is considered by Voorman et al. [85], which is testing a null hypothesis depending on the tunning parameter, and hence differs philosophically from our work. Asymptotically

normal tests were proposed by Fan and Lv [20], Bradic et al. [10], in the low-dimensional regime, relying on oracle properties. Oracle properties require strong conditions, such as the minimal signal condition which cannot be evaluated in practice. In contrast, our work does not rely on oracle properties or variable selection consistency and can work in high-dimensional settings. Moreover, all of the above propositions, rely on the existence of a likelihood, or more generally on the existence of a loss function. As we commented earlier, every sufficiently smooth convex loss function can easily be translated into the estimating equation framework. Hence our procedure provides alternative confidence regions and hypothesis tests, which are optimal and asymptotically equivalent to the ones considered some settings above, in cases where the likelihood function is available and is sufficiently smooth. However, the distinctive feature of our procedure, is that it handles the estimating equation directly, enabling us to perform inference in many examples which could not be addressed with any of the existing methods. For instance in the paper [35], the authors describe a novel procedure for testing and confidence regions for inverse covariance estimation. This procedure, based on the graphical LASSO estimator [23] is inspired by van de Geer et al. [80], while in our case the CLIME procedure immediately falls under the umbrella of the generic framework we are proposing.

4.1.2 Organization of the Chapter

This chapter is organized as follows. In Section 4.2 we briefly review the framework of conventional estimating equations, and summarize our generic testing procedure for high dimensional equations. In Section 4.3, we layout the foundations of the general theoretical framework. Section 4.4 is dedicated to applying the procedure to the Dantzig Selector. In Section 4.5 we study testing in high-dimensional graphical models. Section 4.6 deals with testing parameters in the sparse LDA. Autoregression models are considered in Section 4.7. Section 4.8 discusses Quasi-Likelihood equations with a canonical link function. Numerical studies are presented in Section 4.9, and a discussion is provided in Section 4.10.

4.1.3 Notation

The following notations are used throughout the chapter. For a vector $\mathbf{v} = (v_1, \ldots, v_d)^T \in \mathbb{R}^d$, let $\|\mathbf{v}\|_q = (\sum_{i=1}^d v_i^q)^{1/q}, 1 \leq q < \infty$, $\|\mathbf{v}\|_0 = |\operatorname{supp}(\mathbf{v})|$, where $\operatorname{supp}(\mathbf{v}) = \{j : v_j \neq 0\}$, and $|A|$ denotes the cardinality of a set A. Furthermore let $\|\mathbf{v}\|_\infty = \max_i |v_i|$ and $\mathbf{v}^{\otimes 2} = \mathbf{v}\mathbf{v}^T$. For a matrix $\mathbf{M}$ denote with $\mathbf{M}_{*j}$ and $\mathbf{M}_{j*}$ the j^{th} column and row of $\mathbf{M}$ correspondingly. Furthermore, let $\|\mathbf{M}\|_{\max} = \max_{ij} |M_{ij}|$, $\|\mathbf{M}\|_p = \max_{\|v\|_p=1} \|\mathbf{M}v\|_p$ for $p \geq 1$. If $\mathbf{M}$ is positive semidefinite let $\lambda_{\max}(\mathbf{M})$ and $\lambda_{\min}(\mathbf{M})$ denote the largest and smallest eigenvalues correspondingly. For a set $S \subset \{1, \ldots d\}$ let $\mathbf{v}_S = \{v_j : j \in S\}$ and S^c be the complement of S. We denote with $\phi, \Phi, \overline{\Phi}$ the pdf, cdf and tail probability of a standard normal random variable correspondingly.

Recall that a random variable is called sub-exponential if there exists a constant $K_1 > 0$ such that $\mathbb{P}(|\boldsymbol{X}| > t) \leq \exp(1 - t/K_1)$ for all $t \geq 0$. We denote the sub-exponential norm

$$\|\boldsymbol{X}\|_{\psi_1} = \sup_{p \geq 1} p^{-1} (\mathbb{E}|\boldsymbol{X}|^p)^{1/p}. \tag{4.1.2}$$

Similarly, a random variable is called sub-Gaussian if there exists a $K_2 > 0$ such that $\mathbb{P}(|\boldsymbol{X}| > t) \leq \exp(1 - t^2/K_2^2)$ for all $t \geq 0$. We denote the sub-Gaussian norm

$$\|\boldsymbol{X}\|_{\psi_2} = \sup_{p \geq 1} p^{-1/2} (\mathbb{E}|\boldsymbol{X}|^p)^{1/p}. \tag{4.1.3}$$

Finally we recall a definition taken from Bickel et al. [9], referred to as the *restricted eigenvalue* (RE) assumption.

Definition 4.1.1 (RE). *We say that the symmetric positive semi-definite matrix $\mathbf{M}_{k \times k}$ possesses the restricted eigenvalue property if:*

$$\operatorname{RE}_{\mathbf{M}}(s, \xi) = \min_{S \subset \{1, \ldots, k\}, |S| \leq s} \min_{\mathbf{u}} \left\{ \frac{\mathbf{u}^T \mathbf{M} \mathbf{u}}{\|\mathbf{u}_S\|_2^2} : \mathbf{u} \in \mathbb{R}^d \setminus \{0\}, \|\mathbf{u}_{S^c}\|_1 \leq \xi \|\mathbf{u}_S\|_1 \right\} > 0.$$

4.2 High Dimensional Estimating Equations

In this section we introduce our generic framework and notations. We review basic properties of standard Z estimators conceded with the case when $d < n$ is fixed, and contrast them to Z estimation when the dimension d is high.

4.2.1 Conventional Z Estimation

In this subsection, we briefly review the conventional Z estimators. For a more thorough review, we direct the interested reader to [68,83]. Assume that we observe n iid copies $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ of a q-dimensional random variable $\boldsymbol{X}$. Let $\mathbf{h} : \mathbb{R}^q \times \mathbb{R}^d \mapsto \mathbb{R}^d$ be a vector valued smooth function. The function $\mathbf{h}$ defines the following equation $\mathbb{E}\mathbf{h}(\boldsymbol{X}, \boldsymbol{\beta}) = 0$, and assuming that this equation has a unique solution in the parameter space $\boldsymbol{\beta} \in \boldsymbol{\Omega} \subset \mathbb{R}^d$, it follows that the function $\mathbf{h}$ determines a "true" parameter value which we denote with — $\boldsymbol{\beta}^*$. In the conventional framework, the dimension d is typically held fixed, and the sample size is allowed to diverge to ∞. In order for us to estimate $\boldsymbol{\beta}^*$, it is natural to translate the population version of the equation into the finite sample version:

$$\frac{1}{n}\sum_{i=1}^{n} \mathbf{h}(\boldsymbol{X}_i, \boldsymbol{\beta}) = 0. \tag{4.2.1}$$

When the dimension d is assumed to be fixed, it can be shown, that under certain regularity conditions on $\mathbf{h}$ and $\boldsymbol{X}$ the above equation produces an estimate $\widehat{\boldsymbol{\beta}}$, which is consistent, and asymptotically normal, e.g. see Van der Vaart [83] (Sections 5.3 and 5.4) for more details. Note, that it is essential that the dimension $d \leq n$, as otherwise the finite sample equation can have multiple solutions. Assume that $\mathbf{h}$ is continuously differentiable, and let $\mathbf{H} : \mathbb{R}^q \times \mathbb{R}^d \mapsto \mathbb{R}^d \times \mathbb{R}^d$ be $\mathbf{H}(\boldsymbol{X}, \boldsymbol{\beta}) := \frac{\partial \mathbf{h}(\boldsymbol{X}, \boldsymbol{\beta})}{\boldsymbol{\beta}^T}$. The intuition behind the normality of $\boldsymbol{\beta}$, can be seen along the lines of the

following Taylor expansion:

$$\frac{1}{n^{1/2}}\sum_{i=1}^{n}\mathbf{H}(\boldsymbol{X}_i,\widetilde{\boldsymbol{\beta}})(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}^*)=-\frac{1}{n^{1/2}}\sum_{i=1}^{n}\mathbf{h}(\boldsymbol{X}_i,\boldsymbol{\beta}^*),$$

where $\widetilde{\boldsymbol{\beta}}=v\widehat{\boldsymbol{\beta}}+(1-v)\boldsymbol{\beta}^*$ for some $v\in[0,1]$. Under regularity conditions first term on the RHS can be seen to converge to a normal distribution, and the term $\frac{1}{n}\sum_{i=1}^{n}\mathbf{H}(\boldsymbol{X}_i,\widetilde{\boldsymbol{\beta}})$ can be shown to be consistent for $\mathbb{E}\mathbf{H}(\boldsymbol{X}_i,\boldsymbol{\beta}^*)$. These facts suggest the following weak convergence result:

$$\sqrt{n}(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}^*)\rightsquigarrow N(0,\boldsymbol{\Sigma}),$$

where

$$\boldsymbol{\Sigma}=[\mathbb{E}\mathbf{H}(\boldsymbol{X},\boldsymbol{\beta}^*)]^{-1}\mathbb{E}[\mathbf{h}(\boldsymbol{X},\boldsymbol{\beta}^*)\mathbf{h}^T(\boldsymbol{X},\boldsymbol{\beta}^*)][\mathbb{E}\mathbf{H}(\boldsymbol{X},\boldsymbol{\beta}^*)]^{-1,T}.$$

To this end note that from estimation perspective, in cases when we have a likelihood function available and $\mathbf{h}=\frac{\partial\ell}{\partial\boldsymbol{\beta}}$, such an estimating equation is optimal in the sense that the estimator $\widehat{\boldsymbol{\beta}}$ achieves the Cramer-Rao lower bound and has minimum variance.

The above reasoning not only gives us the intuition behind the normality of the Z estimator — $\widehat{\boldsymbol{\beta}}$, but also suggests a way to test whether certain coordinate of the vector $\boldsymbol{\beta}$ is 0. Let us assume that the vector $\boldsymbol{\beta}=(\theta,\boldsymbol{\gamma}^T)^T$, and we are interested in testing whether the one-dimensional component $H_0:\theta=0$ versus the non-restricted alternative $H_A:\theta\neq 0$. Throughout this chapter, we will assume without loss of generality that θ is the first component of $\boldsymbol{\beta}$. In this scenario the parameters $\boldsymbol{\gamma}$ are nuisance.

One possibility to test in the parameter θ in this situation is to conduct a so-called "Wald" test, relying on the asymptotic distribution of $\sqrt{n}(\widehat{\theta}-0)$. Such a test compares the value of $\sqrt{n}(\widehat{\theta}-0)$ to a quantile of $N(0,\widehat{\sigma}^2)$, where $\widehat{\sigma}^2$ is a consistent estimate for $\sigma^2=\boldsymbol{\Sigma}_{11}$.

To this end, note that the following expression has exactly the same asymptotic distribution as

$\sqrt{n}(\widehat{\theta} - 0)$:

$$n^{-1/2}\mathbf{v}^{*T}\sum_{i=1}^{n}\mathbf{h}(\boldsymbol{X}_i, \boldsymbol{\beta}^*),$$

by the CLT, where $\mathbf{v}^{*T} = [\mathbb{E}\mathbf{H}(\boldsymbol{X}, \boldsymbol{\beta}^*)]^{-1}_{1*}$. The above expression can be viewed as a projection of the estimating equation, evaluated at the true parameter. Hence if we are able to consistently estimate the expression above, we will achieve an asymptotically equivalent test to the Wald test. In the low dimensional framework a natural candidate for such an estimate under the null is $n^{-1/2}\widehat{\mathbf{v}}^T\sum_{i=1}^{n}\mathbf{h}(\boldsymbol{X}_i, \widehat{\boldsymbol{\beta}}_0)$, where $\widehat{\boldsymbol{\beta}}_0 = (0, \widehat{\boldsymbol{\gamma}}^T)^T$ and $\widehat{\mathbf{v}}^T = [n^{-1}\sum_{i=1}^{n}\mathbf{H}(\boldsymbol{X}_i, \widehat{\boldsymbol{\beta}})]^{-1}_{1*}$. Below we consider a natural extension of this framework in the growing d with n case.

4.2.2 High-Dimensional Framework

As we mentioned in Section 4.2.1, in the case when $d > n$, conventional Z estimation fails as one has more parameters than samples. To deal with such situations, we borrow ideas from the Dantzig Selector[14]. Assuming that the underlying true parameter $\boldsymbol{\beta}^*$ is sparse, instead of solving (4.2.1) precisely, we will solve the following optimization problem:

$$\widehat{\boldsymbol{\beta}} = \text{argmin}\, \|\boldsymbol{\beta}\|_1 \;\text{ s.t }\; \left\|\frac{1}{n}\sum_{i=1}^{n}\mathbf{h}(\boldsymbol{X}_i, \boldsymbol{\beta})\right\|_\infty \leq \lambda. \tag{4.2.2}$$

Following the conventional Z estimation approach, we define the following projected test function:

$$\widehat{S}(\boldsymbol{\beta}) = \frac{1}{n}\sum_{i=1}^{n}\widehat{\mathbf{v}}^T\mathbf{h}(\boldsymbol{X}_i, \boldsymbol{\beta}),$$

where the vector $\widehat{\mathbf{v}}$ is defined as the solution to the optimization problem:

$$\widehat{\mathbf{v}} = \text{argmin}\, \|\mathbf{v}\|_1 \;\text{ s.t }\; \left\|\frac{1}{n}\sum_{i=1}^{n}\mathbf{v}^T\mathbf{H}(\boldsymbol{X}_i, \widehat{\boldsymbol{\beta}}) - \mathbf{e}\right\|_\infty \leq \lambda'. \tag{4.2.3}$$

Here, $\mathbf{e}$ is a d-dimensional row vector $(1, 0, \ldots, 0)$, where the position of 1 corresponds to that of θ among $\boldsymbol{\beta}$. Note here that the matrix $\mathbf{H}(\boldsymbol{X}_i, \boldsymbol{\beta}) = (\frac{\partial \mathbf{h}(\boldsymbol{X}_i, \boldsymbol{\beta})}{\partial \beta_1}, \ldots, \frac{\partial \mathbf{h}(\boldsymbol{X}_i, \boldsymbol{\beta})}{\partial \beta_d})$ need not be symmetric in general. The population version of $\widehat{\mathbf{v}}$ is as defined in the previous Section: $\mathbf{v}^* = [\mathbb{E}\mathbf{H}(\boldsymbol{X}, \boldsymbol{\beta}^*)]_{1*}^{-1,T}$. One of the crucial assumptions for the test which we present below to work is that the row vector $\mathbf{v}^{*T}$ and the true parameter $\boldsymbol{\beta}^*$ are sufficiently sparse.

In order to perform a test in the high-dimensional framework, one needs to evaluate $n^{1/2}\widehat{S}(\widehat{\boldsymbol{\beta}}_0)$, and compare the value to a $N(0, \widehat{\sigma}^2)$ random variable, where a consistent estimate $\widehat{\sigma}^2$ of $\sigma^2 = \mathrm{Var}(\mathbf{v}^{*T}\mathbf{h}(\boldsymbol{X}, \boldsymbol{\beta}^*))$ needs to be given. Here $\widehat{\boldsymbol{\beta}}_0 = (0, \widehat{\boldsymbol{\gamma}}^T)^T$, is an estimate of $\boldsymbol{\beta}^*$ under the null hypothesis. We summarize the calculation of the test statistic for high dimensional estimating equations in the following:

Algorithm 4 Test Statistic for High-Dimensional Linear Equations

Input: Data $\{\boldsymbol{X}_i\}_{i=1}^n$, $\mathbf{h}$; Tuning parameters λ, λ',

1. Calculate the optimization problem (4.2.2), to obtain an estimate $\widehat{\boldsymbol{\beta}}$:

$$\widehat{\boldsymbol{\beta}} = \text{argmin } \|\boldsymbol{\beta}\|_1 \text{ s.t } \left\| \frac{1}{n} \sum_{i=1}^{n} \mathbf{h}(\boldsymbol{X}_i, \boldsymbol{\beta}) \right\|_\infty \leq \lambda;$$

2. Calculate the projection direction $\widehat{\mathbf{v}}^T$ through the following optimization based on (4.2.3):

$$\widehat{\mathbf{v}} = \text{argmin } \|\mathbf{v}\|_1 \text{ s.t } \left\| \frac{1}{n} \sum_{i=1}^{n} \mathbf{v}^T \mathbf{H}(\boldsymbol{X}_i, \widehat{\boldsymbol{\beta}}) - \mathbf{e} \right\|_\infty \leq \lambda';$$

3. Output the sparse projected test function:

$$\widehat{S}(\boldsymbol{\beta}) = \frac{1}{n} \widehat{\mathbf{v}}^T \sum_{i=1}^{n} \mathbf{h}(\boldsymbol{X}_i, \boldsymbol{\beta})$$

4.3 General Theoretical Framework

Assume that we observe n iid copies $\boldsymbol{X}_1, ..., \boldsymbol{X}_n$ of $\boldsymbol{X}$. In this section we provide sufficient conditions to guarantee that Algorithm 4 will provide us with a statistic which we can use to test $H_0 : \theta = 0$ vs $H_A : \theta \neq 0$. Our results show that if properly normalized the output statistic from Algorithm 4 will converge weakly to a standard normal random variable. Furthermore, we show how to use our framework to construct confidence intervals.

We will denote with $\mathbb{P}_{\boldsymbol{\beta}}$ the probability measure corresponding to the distribution of $\boldsymbol{X}_i$ generated with a parameter $\boldsymbol{\beta}$. We will use the shorthand notation $\mathbb{P}^* = \mathbb{P}_{\boldsymbol{\beta}^*}$, to indicate the measure

corresponding to the true parameter $\boldsymbol{\beta}^*$.

4.3.1 Weak Convergence under the Null Hypothesis

Below we make several assumptions which are needed to establish the weak convergence. Note that the true parameter under the null distribution has the form $\boldsymbol{\beta}^* = (0, \boldsymbol{\gamma}^{*T})^T$ and the two will be used interchangeably. Denote with $\widehat{\boldsymbol{\beta}}_0 = (0, \widehat{\boldsymbol{\gamma}}^T)^T$, where $\widehat{\boldsymbol{\gamma}}$ is the estimate of nuisance parameter part from Algorithm 4.

Assumption 4.3.1 (Consistent Estimation)**.**

$$\lim_{n\to\infty} \mathbb{P}^*(\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_1 \leq r_1(n)) = 1, \tag{4.3.1}$$

$$\lim_{n\to\infty} \mathbb{P}^*(\|\widehat{\mathbf{v}} - \mathbf{v}^*\|_1 \leq r_2(n)) = 1, \tag{4.3.2}$$

where $r_1(n), r_2(n) = o(1)$.

Assumption 4.3.2 (Noise Condition)**.**

$$\lim_{n\to\infty} \mathbb{P}^*\left(\left\|\frac{1}{n}\sum_{i=1}^{n} \mathbf{h}(\boldsymbol{X}_i, \boldsymbol{\beta}^*)\right\|_\infty \leq r_3(n)\right) = 1 \tag{4.3.3}$$

$$\lim_{n\to\infty} \mathbb{P}^*\left(\sup_{\nu\in[0,1]} \left\|\frac{1}{n}\sum_{i=1}^{n} \widehat{\mathbf{v}}^T \left[\mathbf{H}(\boldsymbol{X}_i, \widetilde{\boldsymbol{\beta}}_\nu)\right]_{-1}\right\|_\infty \leq r_4(n)\right) = 1 \tag{4.3.4}$$

where, by $[\cdot]_{-1}$ *we mean dropping the first column,* $r_3(n), r_4(n) = o(1)$*, and* $\widetilde{\boldsymbol{\beta}}_\nu = \nu\widehat{\boldsymbol{\beta}}_0 + (1-\nu)\boldsymbol{\beta}^*$.

Next we show a theorem which gives us an influence function expansion of our test function S.

Theorem 4.3.3. *Suppose assumptions (4.3.1), (4.3.2), (4.3.3) and (4.3.4) hold in such a way so that*

$$n^{1/2}(r_1(n)r_4(n) + r_2(n)r_3(n)) = o(1) \tag{4.3.5}$$

Then under H_0 we have the following influence function expansion:

$$n^{1/2}\widehat{S}(\widehat{\boldsymbol{\beta}}_0) = n^{1/2}S(\boldsymbol{\beta}^*) + o_p(1) = \frac{1}{n^{1/2}}\sum_{i=1}^{n}\mathbf{v}^{*T}\mathbf{h}(\boldsymbol{X}_i, \boldsymbol{\beta}^*) + o_p(1)$$

The proof of this Theorem can be found in Appendix C.1.

Assumption 4.3.4 (CLT)**.**

$$\frac{1}{(\mathbf{v}^{*T}\boldsymbol{\Sigma}\mathbf{v}^*)^{1/2}n^{1/2}}\sum_{i=1}^{n}\mathbf{v}^{*T}\mathbf{h}(\boldsymbol{X}_i, \boldsymbol{\beta}^*) \rightsquigarrow N(0,1), \tag{4.3.6}$$

where $\boldsymbol{\Sigma} = \operatorname{Cov}\mathbf{h}(\boldsymbol{X}, \boldsymbol{\beta}^*)$, and it is assumed that $\mathbf{v}^{*T}\boldsymbol{\Sigma}\mathbf{v}^* \geq C > 0$

Corollary 4.3.5. *Assume the same assumptions, as in Theorem 4.3.3, and in addition assume condition 4.3.4. We have that:*

$$\frac{n^{1/2}}{\sqrt{\mathbf{v}^{*T}\boldsymbol{\Sigma}\mathbf{v}^*}}\widehat{S}(\widehat{\boldsymbol{\beta}}_0) \rightsquigarrow N(0,1)$$

This Corollary follows immediately from the influence function representation in Theorem 4.3.3 and thus we omit the proof. It is clear that in practice we cannot use the above as a test statistic since we do not know the precise values of $\mathbf{v}^*$ or $\boldsymbol{\Sigma}$. If a consistent estimate of $\mathbf{v}^{*T}\boldsymbol{\Sigma}\mathbf{v}^*$ — $\widehat{\sigma}^2$ is provided, the following is an immediate consequence of Slutsky's theorem:

Proposition 4.3.6. *Assume that $\widehat{\sigma}^2$ is any consistent estimator of $\mathbf{v}^{*T}\boldsymbol{\Sigma}\mathbf{v}^*$. We then have that for $\widehat{U}_n = \frac{n^{1/2}}{\widehat{\sigma}}\widehat{S}(\widehat{\boldsymbol{\beta}}_0)$:*

$$\lim_{n\to\infty}|\mathbb{P}^*(\widehat{U}_n \leq t) - \Phi(t))| = 0.$$

We now provide generic sufficient conditions for constructing such a consistent estimate. Define $\widehat{\boldsymbol{\Sigma}} := \frac{1}{n}\sum_{i=1}^{n}\mathbf{h}(\boldsymbol{X}_i, \widehat{\boldsymbol{\beta}})^{\otimes 2}$. A great candidate for an estimate of $\mathbf{v}^{*T}\boldsymbol{\Sigma}\mathbf{v}^*$ seems to be the "plugin"

estimator: $\widehat{\sigma}^2 = \widehat{\mathbf{v}}^T \widehat{\mathbf{\Sigma}} \widehat{\mathbf{v}}$. Define the statistic: $\widehat{U}_n = \frac{n^{1/2}}{\sqrt{\widehat{\mathbf{v}}^T \widehat{\mathbf{\Sigma}} \widehat{\mathbf{v}}}} \widehat{S}(\widehat{\boldsymbol{\beta}}_0)$. We are interested in showing that $\widehat{U}_n$ converges weakly to a standard normal distribution. To this end we define the following assumption:

Assumption 4.3.7 (Variance Consistency)**.** *Assume that the following holds:*

$$\lim_{n \to \infty} \mathbb{P}^*(\|\widehat{\mathbf{\Sigma}} - \mathbf{\Sigma}\|_{\max} \leq r_5(n)) = 1,$$

where $r_5(n) = o(1)$.

Proposition 4.3.8. *Assume the same assumptions as in Corollary 4.3.5 plus Assumption 4.3.7. Furthermore if we assume that* $\|\mathbf{\Sigma}\|_{\max} = O(1)$, $\|\mathbf{v}^{*T} \mathbf{\Sigma}\|_\infty r_2(n) = o(1)$ *and* $\|\mathbf{v}^*\|_1^2 r_5(n) = o(1)$, *then for any* $t \in \mathbb{R}$ *we have:*

$$\lim_{n \to \infty} |\mathbb{P}^*(\widehat{U}_n \leq t) - \Phi(t))| = 0.$$

The proof of Proposition 4.3.8 can be found in Appendix C.1.

Remark 4.3.9. *Note that Proposition 4.3.8, justifies testing using the statistic* $\widehat{U}_n$. *In other words, testing based on the following rule:*

$$T_n = \begin{cases} 0, \text{ if } |\widehat{U}_n| \leq \Phi^{-1}(1 - \alpha/2), \\ 1, \text{ if } |\widehat{U}_n| > \Phi^{-1}(1 - \alpha/2), \end{cases}$$

where we reject iff $T_n = 1$, *has an asymptotic size* α *under the null.*

4.3.2 Uniform Weak Convergence under the Null

In the previous section we established that if $\boldsymbol{\beta}^* = (0, \boldsymbol{\gamma}^*)$ with $\boldsymbol{\gamma}^*$ being held fixed, the output of Algorithm 4, properly normalized to $\widehat{U}_n$ will have the correct size asymptotically. In this Section we strengthen the assumptions to provide a result which guarantees that the size will be correct uniformly over the following parameter space:

$$\boldsymbol{\Omega}_0 = \{(0, \boldsymbol{\gamma}) : \|\boldsymbol{\gamma}\|_0 \leq s^*\}.$$

We restrict our attention to $\boldsymbol{\Omega}_0$ since we need the parameter $\boldsymbol{\gamma}$ to be sufficiently sparse in order for us to estimate consistently the parameter $\boldsymbol{\beta}$. We now introduce the uniform versions of the assumptions in the preceding Section. Let $\boldsymbol{\beta}_0 = (0, \boldsymbol{\gamma}^T)^T$. Of course when $\boldsymbol{\beta} \in \boldsymbol{\Omega}_0$ we have $\boldsymbol{\beta}_0 \equiv \boldsymbol{\beta}$, but this distinction will become more apparent in the next section.

Assumption 4.3.10 (Uniform Consistent Estimation)**.**

$$\lim_{n\to\infty} \inf_{\beta \in \boldsymbol{\Omega}_0} \mathbb{P}_{\boldsymbol{\beta}}(\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|_1 \leq r_1(n)) = 1, \tag{4.3.7}$$

$$\lim_{n\to\infty} \inf_{\beta \in \boldsymbol{\Omega}_0} \mathbb{P}_{\boldsymbol{\beta}}(\|\widehat{\mathbf{v}} - \mathbf{v}\|_1 \leq r_2(n)) = 1, \tag{4.3.8}$$

where $r_1(n), r_2(n) = o(1)$*. Denote the events* $\mathcal{G}_1^{\boldsymbol{\beta}} = \{\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|_1 \leq r_1(n)\}$*, and* $\mathcal{G}_2^{\boldsymbol{\beta}} = \{\|\widehat{\mathbf{v}} - \mathbf{v}\|_1 \leq r_1(n)\}$

Assumption 4.3.11 (Uniform Noise Condition)**.**

$$\lim_{n\to\infty} \inf_{\boldsymbol{\beta} \in \boldsymbol{\Omega}_0} \mathbb{P}_{\boldsymbol{\beta}}\left(\left\| \frac{1}{n} \sum_{i=1}^{n} \mathbf{h}(\boldsymbol{X}_i, \boldsymbol{\beta}_0) \right\|_\infty \leq r_3(n) \right) = 1 \tag{4.3.9}$$

$$\lim_{n\to\infty} \inf_{\boldsymbol{\beta} \in \boldsymbol{\Omega}_0} \mathbb{P}_{\boldsymbol{\beta}}\left(\sup_{\nu \in [0,1]} \left\| \frac{1}{n} \sum_{i=1}^{n} \widehat{\mathbf{v}}^T \left[\mathbf{H}(\boldsymbol{X}_i, \widetilde{\boldsymbol{\beta}}_\nu) \right]_{-1} \right\|_\infty \leq r_4(n) \right) = 1 \tag{4.3.10}$$

where $r_3(n), r_4(n) = o(1)$ *and* $\widetilde{\boldsymbol{\beta}}_\nu = \nu\widehat{\boldsymbol{\beta}}_0 + (1-\nu)\boldsymbol{\beta}_0$. *Denote with* $\mathcal{G}_3^{\boldsymbol{\beta}} = \left\{ \left\| \frac{1}{n}\sum_{i=1}^n \mathbf{h}(\boldsymbol{X}_i, \boldsymbol{\beta}_0) \right\|_\infty \leq r_3(n) \right\}$ *and* $\mathcal{G}_4^{\boldsymbol{\beta}} = \left\{ \sup_{\nu\in[0,1]} \left\| \frac{1}{n}\sum_{i=1}^n \widehat{\mathbf{v}}^T \left[\mathbf{H}(\boldsymbol{X}_i, \widetilde{\boldsymbol{\beta}}_\nu) \right]_{-1} \right\|_\infty \leq r_4(n) \right\}$.

Assumption 4.3.12 (Uniform CLT)**.**

$$\lim_{n\to\infty} \sup_{\boldsymbol{\beta}\in\boldsymbol{\Omega}_0} \sup_t \left| \mathbb{P}_{\boldsymbol{\beta}} \left(\frac{1}{(\mathbf{v}^T\boldsymbol{\Sigma}\mathbf{v})^{1/2} n^{1/2}} \sum_{i=1}^n \mathbf{v}^T \mathbf{h}(\boldsymbol{X}_i, \boldsymbol{\beta}) \leq t \right) - \Phi(t) \right| = 0 \tag{4.3.11}$$

where $\boldsymbol{\Sigma} = \operatorname{Cov} \mathbf{h}(\boldsymbol{X}, \boldsymbol{\beta})$*, and it is assumed that* $\inf_{\boldsymbol{\beta}\in\boldsymbol{\Omega}_0} \mathbf{v}^T\boldsymbol{\Sigma}\mathbf{v} \geq C > 0$.

Assumption 4.3.13 (Uniform Variance Consistency)**.** *Assume there exists an estimator* $\widehat{\sigma}^2$ *of* $\mathbf{v}^T\boldsymbol{\Sigma}\mathbf{v}$*, such that:*

$$\lim_{n\to\infty} \inf_{\boldsymbol{\beta}\in\boldsymbol{\Omega}_0} \mathbb{P}_{\boldsymbol{\beta}}(|\widehat{\sigma}^2 - \mathbf{v}^T\boldsymbol{\Sigma}\mathbf{v}| \leq \tau(n)) = 1,$$

where $\tau(n) = o(1)$*. Let* $\mathcal{G}_5^{\boldsymbol{\beta}} = \{|\widehat{\sigma}^2 - \mathbf{v}^T\boldsymbol{\Sigma}\mathbf{v}| \leq \tau(n)\}$.

We next formulate a theorem which strengthens Proposition 4.3.6. Its proof can be found in Appendix C.1.

Theorem 4.3.14. *Under Assumptions 4.3.10 — 4.3.13, and the further assume that:*

$$n^{1/2}(r_1(n)r_4(n) + r_2(n)r_3(n)) = o(1),$$

Define $\widehat{U}_n = \frac{n^{1/2}}{\widehat{\sigma}}\widehat{S}(0, \widehat{\boldsymbol{\gamma}})$*. Then we have:*

$$\lim_{n\to\infty} \sup_{\boldsymbol{\beta}\in\boldsymbol{\Omega}_0} \sup_t |\mathbb{P}_{\boldsymbol{\beta}}(\widehat{U}_n \leq t) - \Phi(t))| = 0.$$

Next we provide a sufficient conditions, so that the plugin estimate $\widehat{\sigma}^2 = \widehat{\mathbf{v}}^T\widehat{\boldsymbol{\Sigma}}\widehat{\mathbf{v}}$ satisfies assumption 4.3.13.

Assumption 4.3.15 (Plugin Variance Consistency). *Assume that the following holds:*

$$\lim_{n\to\infty} \inf_{\boldsymbol{\beta}\in\boldsymbol{\Omega}_0} \mathbb{P}_{\boldsymbol{\beta}}(\|\widehat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}\|_{\max} \leq r_6(n)) = 1,$$

where $r_6(n) = o(1)$.

We then have the following:

Theorem 4.3.16. *Under Assumptions 4.3.10 — 4.3.12 and 4.3.15, and the further assume that* $\sup_{\boldsymbol{\beta}\in\boldsymbol{\Omega}_0} \|\boldsymbol{\Sigma}\|_{\max} = O(1)$, $\sup_{\boldsymbol{\beta}\in\boldsymbol{\Omega}_0} \|\mathbf{v}^T\boldsymbol{\Sigma}\|_\infty r_2(n) = o(1)$ *and* $\sup_{\boldsymbol{\beta}\in\boldsymbol{\Omega}_0} \|\mathbf{v}\|_1^2 r_6(n) = o(1)$*:*

$$n^{1/2}(r_1(n)r_4(n) + r_2(n)r_3(n)) = o(1),$$

we have:

$$\lim_{n\to\infty} \sup_{\boldsymbol{\beta}\in\boldsymbol{\Omega}_0} \sup_t |\mathbb{P}_{\boldsymbol{\beta}}(\widehat{U}_n \leq t) - \Phi(t))| = 0.$$

Theorem 4.3.16 is provided without proof, as it follows from Theorem 4.3.14, upon recognizing that under the sufficient conditions the plugin estimate $\widehat{\sigma}^2 = \widehat{\mathbf{v}}^T\widehat{\boldsymbol{\Sigma}}\widehat{\mathbf{v}}$ satisfies Assumption 4.3.13.

The assumptions we consider in this Section are clearly stronger than the corresponding assumptions in Section 4.3.1. The reason for strengthening these conditions, is in order to show uniform convergence in contrast to the weak convergence provided in Proposition 4.3.8.

4.3.3 Local Power

In this section we analyze the power of the suggested test with respect to a sequence of local alternatives. To this end we define the following parameter space, which is of interest:

$$\boldsymbol{\Omega}_1(K, \phi) := \{(\theta, \boldsymbol{\gamma}) : \theta = Kn^{-\phi}, \|\boldsymbol{\gamma}\|_0 \leq s^*\},$$

where $s^* = \|\boldsymbol{\gamma}^*\|_0$, and $\phi > 0$ is a parameter determining how fast the alternative is approaching the null. Note that intuitively, as ϕ grows it will become harder to distinguish the alternative from the null. Next we define assumptions in analogue to the one in the previous Section. All events $\mathcal{G}_i^{\boldsymbol{\beta}}, i = 1, \ldots, 5$ are defined as in Section 4.3.2.

Assumption 4.3.17 (Uniform Consistent Estimation)**.**

$$\lim_{n\to\infty} \inf_{\beta \in \boldsymbol{\Omega}_1(K,\phi)} \mathbb{P}_{\boldsymbol{\beta}}(\mathcal{G}_1^{\boldsymbol{\beta}}) = 1, \tag{4.3.12}$$

$$\lim_{n\to\infty} \inf_{\beta \in \boldsymbol{\Omega}_1(K,\phi)} \mathbb{P}_{\boldsymbol{\beta}}(\mathcal{G}_2^{\boldsymbol{\beta}}) = 1, \tag{4.3.13}$$

where $r_1(n), r_2(n) = o(1)$.

Assumption 4.3.18 (Uniform Noise Condition)**.**

$$\lim_{n\to\infty} \inf_{\boldsymbol{\beta} \in \boldsymbol{\Omega}_1(K,\phi)} \mathbb{P}_{\boldsymbol{\beta}}(\mathcal{G}_3^{\boldsymbol{\beta}}) = 1 \tag{4.3.14}$$

$$\lim_{n\to\infty} \inf_{\boldsymbol{\beta} \in \boldsymbol{\Omega}_1(K,\phi)} \mathbb{P}_{\boldsymbol{\beta}}(\mathcal{G}_4^{\boldsymbol{\beta}}) = 1 \tag{4.3.15}$$

where $r_3(n), r_4(n) = o(1)$.

Assumption 4.3.19 (Uniform CLT)**.**

$$\lim_{n\to\infty} \sup_{\boldsymbol{\beta} \in \boldsymbol{\Omega}_1(K,\phi)} \sup_t \left| \mathbb{P}_{\boldsymbol{\beta}}\left(\frac{1}{(\mathbf{v}^T \boldsymbol{\Sigma} \mathbf{v})^{1/2} n^{1/2}} \sum_{i=1}^n \mathbf{v}^T \mathbf{h}(\boldsymbol{X}_i, \boldsymbol{\beta}) \le t \right) - \Phi(t) \right| = 0 \tag{4.3.16}$$

where $\boldsymbol{\Sigma} = \operatorname{Cov} \mathbf{h}(\boldsymbol{X}, \boldsymbol{\beta})$*, and it is assumed that* $\inf_{\boldsymbol{\beta} \in \boldsymbol{\Omega}_1(K,\phi)} \mathbf{v}^T \boldsymbol{\Sigma} \mathbf{v} \ge C > 0$.

Assumption 4.3.20 (Uniform Variance Consistency)**.** *Assume that the following holds:*

$$\lim_{n\to\infty} \inf_{\boldsymbol{\beta} \in \boldsymbol{\Omega}_1(K,\phi)} \mathbb{P}_{\boldsymbol{\beta}}(\mathcal{G}_5^{\boldsymbol{\beta}}) = 1,$$

where $\tau(n) = o(1)$.

Assumption 4.3.21 (Uniform Local Approximation)**.** *Assume that the following holds:*

$$\lim_{n\to\infty} \inf_{\boldsymbol{\beta}\in\boldsymbol{\Omega}_1(K,\phi)} \mathbb{P}_{\boldsymbol{\beta}}(\sqrt{n}|S(\theta,\boldsymbol{\gamma}) - S(0,\boldsymbol{\gamma}) - \theta| \leq r_6(n)) = 1,$$

where $r_6(n) = o(1)$. *We denote with* $\mathcal{G}_6^{\boldsymbol{\beta}} = \{\sqrt{n}|S(\theta,\boldsymbol{\gamma}) - S(0,\boldsymbol{\gamma}) - \theta| \leq r_6(n)\}$.

We are now in position to formulate a theorem for the local power.

Theorem 4.3.22. *Assume that the Assumptions 4.3.17 — 4.3.21 hold and that furthermore we have* $n^{1/2}(r_1(n)r_4(n) + r_2(n)r_3(n)) = o(1)$. *Define* $\widehat{U}_n$ *as in Theorem 4.3.14. Then we have*

$$\lim_{n\to\infty} \sup_{\boldsymbol{\beta}\in\boldsymbol{\Omega}_1(K,\phi)} \sup_t \left|\mathbb{P}_{\boldsymbol{\beta}}(\widehat{U}_n \leq t) - \Phi(t)\right| = 0, \text{ if } \phi > 1/2 \tag{4.3.17}$$

$$\lim_{n\to\infty} \sup_{\boldsymbol{\beta}\in\boldsymbol{\Omega}_1(K,\phi)} \sup_t \left|\mathbb{P}_{\boldsymbol{\beta}}(\widehat{U}_n \leq t) - \Phi\left(t + \frac{K}{\sqrt{\mathbf{v}^T\boldsymbol{\Sigma}\mathbf{v}}}\right)\right| = 0, \text{ if } \phi = 1/2 \tag{4.3.18}$$

and for a fixed $t \in \mathbb{R}$ *and* $K \neq 0$ *we have:*

$$\lim_{n\to\infty} \sup_{\boldsymbol{\beta}\in\boldsymbol{\Omega}_1(K,\phi)} \mathbb{P}_{\boldsymbol{\beta}}(|\widehat{U}_n| \leq t) = 0, \text{ if } \phi < 1/2 \tag{4.3.19}$$

Below we provide a sufficient condition to obtain a consistent estimate of $\mathbf{v}^T\boldsymbol{\Sigma}\mathbf{v}$.

Proposition 4.3.23 (Uniform Plugin Variance Consistency)**.** *Assume that the following holds:*

$$\lim_{n\to\infty} \inf_{\boldsymbol{\beta}\in\boldsymbol{\Omega}_1(K,\phi)} \mathbb{P}_{\boldsymbol{\beta}}(\|\widehat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}^*\|_{\max} \leq r_7(n)) = 1,$$

where $r_7(n) = o(1)$. *Furthermore, assume that Assumptions 4.3.17 — 4.3.19 hold, and in addition we have* $\sup_{\boldsymbol{\beta}\in\boldsymbol{\Omega}_0} \|\boldsymbol{\Sigma}\|_{\max} = O(1)$, $\sup_{\boldsymbol{\beta}\in\boldsymbol{\Omega}_0} \|\mathbf{v}^T\boldsymbol{\Sigma}\|_\infty r_2(n) = o(1)$ *and* $\sup_{\boldsymbol{\beta}\in\boldsymbol{\Omega}_0} \|\mathbf{v}\|_1^2 r_7(n) =$

$o(1)$ *and*

$$n^{1/2}(r_1(n)r_4(n) + r_2(n)r_3(n)) = o(1).$$

Then $\widehat{\sigma}^2 = \widehat{\mathbf{v}}^T\widehat{\mathbf{\Sigma}}\widehat{\mathbf{v}}$ *satisfies Assumption 4.3.20.*

The proof of Proposition 4.3.23 is similar to the proof of Proposition 4.3.8 and we omit it.

4.3.4 One-Step estimator and Confidence Intervals

Next we consider, an approach which will allow us to construct confidence intervals for a parameter of interest — θ. Note that in general, the estimate $\widehat{\theta}$ of θ cannot be expected to be regular, and hence weak convergence to a normal distribution cannot be guaranteed. We can make usage of a principle known as a one-step estimator (see Van der Vaart [83]), to define a modified version of $\widehat{\theta}$ which achieves asymptotic normality. The intuition behind this estimator is based on a Taylor expansion of the test statistic, about a "nice" estimator:

$$\widehat{S}(\theta, \widehat{\gamma}) \approx \widehat{S}(\widehat{\theta}, \widehat{\gamma}) + \frac{\partial}{\partial\theta}\widehat{S}(\widehat{\theta}, \widehat{\gamma})(\theta - \widehat{\theta}).$$

If we could treat $\widehat{S}(\theta, \widehat{\gamma})$ as 0 the above expansion would give raise to the following estimator:

$$\widetilde{\theta} = \widehat{\theta} - \left(\frac{\partial}{\partial\theta}\widehat{S}(\widehat{\theta}, \widehat{\gamma})\right)^{-1}\widehat{S}(\widehat{\theta}, \widehat{\gamma}) \tag{4.3.20}$$

The above estimator has the following explicit form (assuming that θ is located at first position of the vector $\boldsymbol{\beta}$):

$$\widetilde{\theta} = \widehat{\theta} - \frac{\sum_{i=1}^{n}\widehat{\mathbf{v}}^T\mathbf{h}(\boldsymbol{X}_i, \widehat{\boldsymbol{\beta}})}{\sum_{i=1}^{n}\widehat{\mathbf{v}}^T\mathbf{H}(\boldsymbol{X}_i, \widehat{\boldsymbol{\beta}})_{*1}} \tag{4.3.21}$$

Below we formulate several sufficient conditions needed to establish the normality of $\widetilde{\theta}$. Let $\boldsymbol{\beta}_{\theta^*} = (\theta^*, \widehat{\boldsymbol{\gamma}}^T)^T$.

Assumption 4.3.24 (Noise Condition). *Assume that*

$$\lim_{n\to\infty} \mathbb{P}^*\left(\left\|\frac{1}{n}\sum_{i=1}^{n} \mathbf{h}(\boldsymbol{X}_i, \boldsymbol{\beta}^*)\right\|_\infty \leq r_3(n)\right) = 1, \tag{4.3.22}$$

$$\lim_{n\to\infty} \mathbb{P}^*\left(\sup_{\nu\in[0,1]} \left\|\frac{1}{n}\sum_{i=1}^{n} \widehat{\mathbf{v}}^T \left[\mathbf{H}(\boldsymbol{X}_i, \widetilde{\boldsymbol{\beta}}_\nu)\right]_{-1} \right\|_\infty \leq r_4(n)\right) = 1, \tag{4.3.23}$$

where $r_3(n), r_4(n) = o(1)$, *and* $\widetilde{\boldsymbol{\beta}}_\nu = \nu\widehat{\boldsymbol{\beta}}_{\theta^*} + (1-\nu)\boldsymbol{\beta}^*$.

Assumption 4.3.25 (Stability). *Assume that:*

$$\lim_{n\to\infty} \mathbb{P}^*\left(\sup_{\nu\in[0,1]} \left|\frac{1}{n}\sum_{i=1}^{n} \widehat{\mathbf{v}}^T \left[\mathbf{H}(\boldsymbol{X}_i, \widetilde{\boldsymbol{\beta}}_\nu)\right]_{*1} - 1\right| \leq r_5(n)\right) = 1,$$

where $r_5(n) = o(1)$, *and* $\widetilde{\boldsymbol{\beta}}_\nu = \nu\widehat{\boldsymbol{\beta}} + (1-\nu)\widehat{\boldsymbol{\beta}}_{\theta^*}$.

We are now ready to identify the asymptotic distribution of $n^{1/2}(\widetilde{\theta} - \theta^*)$.

Proposition 4.3.26. *Assume that Assumptions 4.3.1, 4.3.4, 4.3.24 and 4.3.25 hold. Assume furthermore that* $n^{1/2}(r_1(n)r_4(n) + r_2(n)r_3(n)) = o(1)$ *and* $n^{1/2}|\widehat{\theta} - \theta^*|r_5(n) = o_p(1)$. *Then we have:*

$$\frac{n^{1/2}}{\sqrt{\mathbf{v}^{*T}\boldsymbol{\Sigma}\mathbf{v}^*}}(\widetilde{\theta} - \theta^*) \rightsquigarrow N(0,1)$$

The proof of the Proposition can be found in Appendix C.1.

Remark 4.3.27. *As it becomes evident from Corollary 4.3.5 and Proposition 4.3.26 the asymptotic distributions of the suggested test and the test based on the one-step estimator are asymptotically equivalent.*

Remark 4.3.28. *Observe that in cases when the estimating equation comes from a log-likelihood, i.e.* $\mathbf{h} = \frac{\partial \ell}{\partial \boldsymbol{\beta}}$*, under regularity conditions we have* $\mathbf{v}^{*T}\boldsymbol{\Sigma}\mathbf{v}^* = (\boldsymbol{\Sigma}^{-1})_{11}$*, since in this case the expected information equals* $-\mathbb{E}\mathbf{H}(\boldsymbol{X}, \boldsymbol{\beta}^*) = \operatorname{Cov}\mathbf{h}(\boldsymbol{X}, \boldsymbol{\beta}^*)$*. In such situations the score equations lead to efficient estimators*[27]*, and the variance* $(\boldsymbol{\Sigma}^{-1})_{11}$ *coincides with the optimal one, hence our estimator is optimal.*

Remark 4.3.29. *Assuming that there exists a consistent estimator* $\widehat{\sigma}^2$ *for* $\mathbf{v}^{*T}\boldsymbol{\Sigma}\mathbf{v}^*$*, it is evident that we can construct* α*-level confidence intervals of* θ^* *of the form* $\widetilde{\theta} \pm \Phi^{-1}(1-\alpha/2)\widehat{\sigma}/\sqrt{n}$*.*

4.3.5 Simplifications for Linear Estimating Equations

In the remaining of this section, we construct Z estimators, based on estimating equations with $\mathbf{h}(\boldsymbol{X}, \boldsymbol{\beta}) = A_b(\boldsymbol{X})\boldsymbol{\beta} - B_b(\boldsymbol{X})$, where $A_b : \mathbb{R}^q \mapsto \mathbb{R}^{d\times d}$, and $B_b : \mathbb{R}^q \mapsto \mathbb{R}^d$, are some deterministic functions.

Assuming that $(\mathbb{E}A_b(\boldsymbol{X}))^{-1}$ exists that the true parameter $\boldsymbol{\beta}$ is defined through:

$$\boldsymbol{\beta}^* = (\mathbb{E}A_b(\boldsymbol{X}))^{-1}(\mathbb{E}B_b(\boldsymbol{X})),$$

in this linear case. The last assumption is equivalent to assuming uniqueness of $\boldsymbol{\beta}^*$ when the parameter space $\boldsymbol{\Omega} = \mathbb{R}^d$. Moreover in this framework $\mathbf{v}^{*T} = (\mathbb{E}A_b(\boldsymbol{X}))^{-1}_{1*}$.

Note that one of the significant simplifications in the linear case is that the function $\mathbf{H}(\boldsymbol{X}, \boldsymbol{\beta}) = A_b(\boldsymbol{X})$, and hence does not depend on the parameter $\widetilde{\boldsymbol{\beta}}$. One immediate implication of this fact is that conditions (4.3.4), (4.3.23) (4.3.10) and (4.3.14) reduce to corresponding assumptions on λ'. In particular (4.3.4) and (4.3.23) can equivalently be expressed as $\lambda' = O(r_4(n))$, (4.3.10) can be stated as $\sup_{\boldsymbol{\beta}\in\boldsymbol{\Omega}_0}\lambda' = O(r_4(n))$ and (4.3.14) is equivalent to $\sup_{\boldsymbol{\beta}\in\boldsymbol{\Omega}_1(K,\phi)}\lambda' = O(r_4(n))$.

Furthermore, in Assumption 4.3.25, we have:

$$\left|\frac{1}{n}\sum_{i=1}^{n}\widehat{\mathbf{v}}^T A_b(\boldsymbol{X}_i)_{*1} - 1\right| \leq \lambda',$$

and thus we can express this assumption as $\lambda' = O(r_5(n))$. Furthermore, note that in the proof of Proposition 4.3.26, the term $I_1 = 0$, and hence the condition $n^{1/2}|\widehat{\theta} - \theta^*|r_5(n) = o_p(1)$ is not necessary in the linear case.

4.4 Dantzig Selector

In this section we will consider an application of the theory developed in Section 4.3 to the linear model. Assume that we have n iid draws from the usual linear regression model with:

$$y = \boldsymbol{X}^T\boldsymbol{\beta}^* + \varepsilon = \boldsymbol{X}_1\theta^* + \boldsymbol{X}_{-1}^T\boldsymbol{\gamma}^* + \varepsilon$$

where ε, is a random variable with $\mathbb{E}(\varepsilon) = 0$ and $\mathrm{Var}(\varepsilon) \geq C_\varepsilon > 0$. We assume that ε is sub-Gaussian, with $\|\varepsilon\|_{\psi_2} = K$. Note that the last implies that $\mathrm{Var}(\varepsilon) \leq 2K^2$. We further assume that each of the coordinates of the vectors $\boldsymbol{X}$ are sampled from a sub-Gaussian distribution, or in other words we are assuming that $K_{\boldsymbol{X}} = \sup_{j\in\{1,\ldots,d\}} \|\boldsymbol{X}^j\|_{\psi_2} < \infty$. Note that throughout we will consider $K_{\boldsymbol{X}}$ as a fixed constant regardless of the increasing dimension d. Furthermore we are assuming that $\boldsymbol{X}$ is sampled independently of the error ϵ. In addition we denote the second moment matrix of $\boldsymbol{X}$ with $\boldsymbol{\Sigma}_{\boldsymbol{X}} = \mathbb{E}\boldsymbol{X}^{\otimes 2}$ and assume that $\boldsymbol{\Sigma}_{\boldsymbol{X}} > \delta$, where $\delta > 0$ is a fixed constant regardless of the increasing dimension d. As before we observe n iid samples from the linear regression $(Y_i, \boldsymbol{X}_i)_{i=1}^n$. Assuming that $\boldsymbol{\beta} = (\theta, \boldsymbol{\gamma})$, we are concerned with testing whether the first component of $\boldsymbol{\beta}$ is zero, i.e. we are concerned with testing $H_0 : \theta = 0$ vs $H_A : \theta \neq 0$.

As mentioned in Section 4.3.5, the true parameter $\boldsymbol{\beta}^*$ clearly solves the d equations: $\mathbb{E}(\mathbf{h}((y, X), \boldsymbol{\beta})) =$

0 when $\mathbf{h}((y, \boldsymbol{X}), \boldsymbol{\beta}) = \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{\beta} - y)$. In other words we have $A_b((y, \boldsymbol{X})) = \boldsymbol{X}^{\otimes 2}$ and $B_b((y, \boldsymbol{X})) = y\boldsymbol{X}$. Thus in the linear regression case, $\widehat{S}(\boldsymbol{\beta})$ reduces to

$$\widehat{S}(\boldsymbol{\beta}) = \frac{1}{n}\sum_{i=1}^{n} \widehat{\mathbf{v}}^T \boldsymbol{X}_i(\boldsymbol{X}_i^T\boldsymbol{\beta} - Y_i),$$

where

$$\widehat{\mathbf{v}} = \text{argmin}\, \|\mathbf{v}\|_1 \text{ s.t } \left\| \frac{1}{n}\sum_{i=1}^{n} \mathbf{v}^T \boldsymbol{X}_i^{\otimes 2} - \mathbf{e} \right\|_{\max} \leq \lambda'.$$

Recall that the population version $\mathbf{v}^*$ is defined through equation (4.2.3), which in this case reduces to:

$$\mathbf{v}^* = \boldsymbol{\Sigma}_{\boldsymbol{X}}^{-1}\mathbf{e}^T.$$

It is essential for our test statistic to produce asymptotically normal results, that the vectors $\mathbf{v}^*$ and $\boldsymbol{\beta}^*$ are sparse. Denote with s and $s_{\mathbf{v}}$ the sparsities of the vectors $\boldsymbol{\beta}^*$ and $\mathbf{v}^*$ correspondingly. Next we proceed to formulate a theorem on the asymptotic normality of $n^{1/2}\widehat{S}(\boldsymbol{\beta})$ under the null hypothesis $\theta = 0$.

Theorem 4.4.1. *Assume that the noise distribution is sub-Gaussian, the covariate distribution is sub-Gaussian with ψ_2 norms as specified above. Furthermore, assume that the smallest eigenvalue of the second moment matrix $\lambda_{\min}(\boldsymbol{\Sigma}_{\boldsymbol{X}}) > \delta > 0$ is bounded away from 0. Let $\|\boldsymbol{\beta}^*\|_0 = s$ and $\|\mathbf{v}^*\|_0 = s_{\mathbf{v}}$. Under the assumption that $\max(s_{\mathbf{v}}, s)\|\mathbf{v}^*\|_1 \frac{\log d}{\sqrt{n}} = o(1)$, and large enough tuning parameters with $\lambda \asymp \sqrt{\frac{\log d}{n}}$ and $\lambda' \asymp \|\mathbf{v}^*\|_1\sqrt{\frac{\log d}{n}}$, we have the following asymptotic influence function expansion of the test statistic:*

$$n^{1/2}\widehat{S}(0, \widehat{\boldsymbol{\gamma}}) = \frac{1}{n^{1/2}}\sum_{i=1}^{n} \mathbf{v}^{*T}\boldsymbol{X}_i(\boldsymbol{X}_{i,-1}^T\boldsymbol{\gamma}^* - Y_i) + o_p(1).$$

Remark 4.4.2. *Note here that it is implied that $\lambda' = o(1)$ and hence since $\|\mathbf{v}^*\|_1 \geq 2K_{\boldsymbol{X}}^{-2}$ it*

follows that $\lambda = o(1)$ *as well.*

Remark 4.4.3. *Observe that* $\|\mathbf{v}^*\|_1 \leq \sqrt{s_{\mathbf{v}}}\|\mathbf{v}^*\|_2 \leq \sqrt{s_{\mathbf{v}}}\delta$*, as we verify in Remark 4.4.4. This yields sufficient conditions by substituting* $\|\mathbf{v}^*\|_1$ *with* $\sqrt{s_{\mathbf{v}}}$*. We note moreover, that under the assumption that* $\mathbf{v}^{*T}\boldsymbol{X}$ *is sub-Gaussian, we can further relax the requirements on sparsity* $s_{\mathbf{v}}$ *dimension* d *and number of observations* n*.*

The proof of this Theorem can be found in Appendix C.2, as the rest of the proofs from this Section. Denote with $\Delta := \mathbf{v}^{*T}\boldsymbol{\Sigma}_{\boldsymbol{X}}\mathbf{v}^* \operatorname{Var}(\varepsilon)$.

Remark 4.4.4. *We note that* $\Delta \geq (\boldsymbol{\Sigma}_{\boldsymbol{X}}^{-1})_{11}C_\varepsilon \geq (\boldsymbol{\Sigma}_{\boldsymbol{X},11})^{-1}C_\varepsilon \geq \frac{C_\varepsilon}{2K_{\boldsymbol{X}}^2} > 0$*. Furthermore observe that* $(\boldsymbol{\Sigma}_{\boldsymbol{X}}^{-1})_{11} = \mathbf{v}^{*T}\boldsymbol{\Sigma}_{\boldsymbol{X}}\mathbf{v}^* \geq \delta\|\mathbf{v}^*\|_2^2 \geq \delta(\boldsymbol{\Sigma}_{\boldsymbol{X}}^{-1})_{11}^2$*. Hence* $\Delta \leq 2K^2\delta^{-1}$*.*

Next we have the following:

Corollary 4.4.5. *Under the same assumptions as Theorem 4.4.1, and the additional assumptions* $\frac{s_{\mathbf{v}}^{3/2}}{n^{1/2}} = o(1)$*, we have that:*

$$\frac{n^{1/2}}{\sqrt{\Delta}}\widehat{S}(0,\widehat{\boldsymbol{\gamma}}) \rightsquigarrow N(0,1).$$

In order for the test we developed above to be applicable in practice, we further need to find consistent estimators for Δ. The proposition below provides us with a consistent estimator of Δ, and thus enables testing in practical settings.

Proposition 4.4.6. *Let*

$$\widehat{\Delta}_1 := \frac{1}{n}\sum_{i=1}^{n}(\widehat{\mathbf{v}}^T\boldsymbol{X}_i)^2\frac{1}{n}\sum_{i=1}^{n}(Y_i - \boldsymbol{X}_i^T\widehat{\boldsymbol{\beta}})^2.$$

Under the assumptions of Theorem 4.4.1, and the following additional assumption:

$$\|\mathbf{v}^*\|_1^2\sqrt{\frac{\log d}{n}} = o(1),$$

we have that:

$$\widehat{\Delta}_1 \to_p \Delta.$$

Remark 4.4.7. *Let* $\widehat{\Delta}_2 = \widehat{\mathbf{v}}_1 n^{-1} \sum_{i=1}^n (Y_i - \boldsymbol{X}_i^T \widehat{\boldsymbol{\beta}})^2$. *Under the assumptions of Theorem 4.4.1, we have* $\widehat{\Delta}_1 \to_p \Delta$.

We furthermore suggest an alternative plug-in estimator in the following:

Proposition 4.4.8. *Under the assumptions of Theorem 4.4.1, and the following additional assumptions:*

$$s^2 \frac{\log d}{n} \log(nd) \|\mathbf{v}^*\|_1 = o(1),$$
$$\|\mathbf{v}^*\|_1^2 \frac{\log(nd)}{\sqrt{n}} = o(1),$$

we have that:

$$\widehat{\Delta}_3 = \frac{1}{n} \sum_{i=1}^n (\widehat{\mathbf{v}}^T \boldsymbol{X}_i (\boldsymbol{X}_i^T \widehat{\boldsymbol{\beta}} - Y_i))^2 \to_p \Delta.$$

Remark 4.4.9. *Note that the assumption of the estimator in Proposition 4.4.8 are slightly stronger than the assumptions in Proposition 4.4.6 which in turn are stronger than the ones in Remark 4.4.7.*

The last propositions suggest two estimates which satisfy the condition in Proposition 4.3.6, and thus the two statistics based on them will provide results with correct size under the null distribution. Hence we can state the following:

Theorem 4.4.10. *Assume all assumptions in Corollary 4.4.5, and construct estimates of* $\Delta - \widehat{\Delta}_1$, $\widehat{\Delta}_2$ *and* $\widehat{\Delta}_3$ *based on Proposition 4.4.6, Remark 4.4.7 or Proposition 4.4.8 under their corresponding conditions. Then the statistics* $\widehat{U}_n^i = \frac{n^{1/2}}{\widehat{\Delta}_i} \widehat{S}(0, \widehat{\boldsymbol{\gamma}})$ *satisfy:*

$$\lim_{n \to \infty} |\mathbb{P}^*(\widehat{U}_n^i \leq t) - \Phi(t))| = 0, \;\; i = 1, 2, 3.$$

Remark 4.4.11. *In fact, carefully inspecting the proof of Theorem 4.4.10, shows that the uniform assumptions from Section 4.3.2 are satisfied, and hence under the same assumptions as in Theorem 4.4.10, we have:*

$$\lim_{n\to\infty} \sup_{\boldsymbol{\beta}\in\boldsymbol{\Omega}_0} \sup_t |\mathbb{P}_{\boldsymbol{\beta}}(\widehat{U}_n^i \leq t) - \Phi(t))| = 0, \;\; i = 1, 2, 3,$$

where $\boldsymbol{\Omega}_0 = \{(0, \boldsymbol{\gamma}^T)^T : \|\boldsymbol{\gamma}\|_0 \leq s\}$.

Finally to conclude this section we present a result on the local power of the proposed test. Recall the definition of the parameters space:

$$\boldsymbol{\Omega}_1(K, \phi) = \{(\theta, \boldsymbol{\gamma}^T)^T : \theta = Kn^{-\phi}, \|\boldsymbol{\gamma}\|_0 \leq s\}$$

Theorem 4.4.12. *Under the same assumptions as in Theorem 4.4.10, and* $Kn^{-\phi}\|\mathbf{v}^*\|_1 s_{\mathbf{v}}\sqrt{\log(d)} = o(1)$*, we have that for* $i = 1, 2, 3$*:*

$$\lim_{n\to\infty} \sup_{\boldsymbol{\beta}\in\boldsymbol{\Omega}_1(K,\phi)} \sup_t \left|\mathbb{P}_{\boldsymbol{\beta}}(\widehat{U}_n^i \leq t) - \Phi(t)\right| = 0, \; \textit{if}\, \phi > 1/2,$$
$$\lim_{n\to\infty} \sup_{\boldsymbol{\beta}\in\boldsymbol{\Omega}_1(K,\phi)} \sup_t \left|\mathbb{P}_{\boldsymbol{\beta}}(\widehat{U}_n^i \leq t) - \Phi\left(t + \frac{K}{\sqrt{\Delta}}\right)\right| = 0, \; \textit{if}\, \phi = 1/2,$$

and for a fixed $t \in \mathbb{R}$ *and* $K \neq 0$ *we have:*

$$\lim_{n\to\infty} \sup_{\boldsymbol{\beta}\in\boldsymbol{\Omega}_1(K,\phi)} \mathbb{P}_{\boldsymbol{\beta}}(|\widehat{U}_n^i| \leq t) = 0, \; \textit{if}\, \phi < 1/2.$$

The proofs of Theorem 4.4.12 can be found in Appendix C.2.

4.4.1 One-Step estimator and Confidence Intervals

Following Section 4.3.4 it's easy to see that the one-step estimate in Section 4.3.4, the Dantzig selector case takes the following form:

$$\widetilde{\theta} = \widehat{\theta} - \frac{\sum_{i=1}^{n} \widehat{\mathbf{v}}^T \boldsymbol{X}_i (\boldsymbol{X}_i^T \widehat{\boldsymbol{\beta}} - Y_i)}{\sum_{i=1}^{n} \widehat{\mathbf{v}}^T \boldsymbol{X}_i \boldsymbol{X}_{i,1}} \tag{4.4.1}$$

We proceed to identify the asymptotic distribution of $n^{1/2}(\widetilde{\theta} - \theta^*)$.

Corollary 4.4.13. *Assume the same assumptions as in Corollary 4.4.5. We then have:*

$$\frac{\sqrt{n}}{\sqrt{\Delta}}(\widetilde{\theta} - \theta^*) \rightsquigarrow N(0, 1)$$

Proof. Using Proposition 4.3.26 it is sufficient to prove that $\lambda' = o_p(1)$. However as we argued in Lemma C.2.5, we can select $\lambda' = C\|\mathbf{v}^*\|_1 \sqrt{\frac{\log d}{n}} = o(1)$, for a large enough C, and this finishes the proof. □

Remark 4.4.14. *As it becomes evident from Corollaries 4.4.5 and 4.4.13 the asymptotic distributions of our proposed test and the one-step estimator based test are asymptotically equivalent. As mentioned in Remark 4.3.28 this estimator is also optimal. Moreover in this case this also implies that the estimator is semi-parametrically efficient[83].*

Remark 4.4.15. *Clearly we can use the plugin estimators suggested in Proposition 4.4.6, Remark 4.4.7 and Proposition 4.4.8, to construct confidence intervals, or test the parameter using the one-step estimator approach suggested above.*

4.5 Edge Testing in Graphical Models

There are many existing procedures for graphical models such as the neighboring pursuit[62], graphical LASSO[95,23], graphical Dantzig Selector[94] and CLIME[13] among others. The majority of the literature focuses on estimation rather than inference with a few recent exceptions e.g.[35]. In this section we consider applications of our general procedures described in Sections 4.2 and 4.3 to Graphical Modeling and inverse covariance estimation.

4.5.1 CLIME

We first turn our attention to the CLIME estimator, suggested in the paper by Cai et al.[13]. We briefly recall the setup here. Assume that $\boldsymbol{X}_1, ..., \boldsymbol{X}_n$ are iid copies of $\boldsymbol{X}$ with $\mathbb{E}(\boldsymbol{X}) = 0$ and $\mathrm{Cov}(\boldsymbol{X}) = \boldsymbol{\Sigma}_{\boldsymbol{X}}$. Denote $\boldsymbol{\Omega}^* = (\boldsymbol{\Sigma}_{\boldsymbol{X}})^{-1}$. When $\boldsymbol{X}_i$ are coming from a Gaussian distribution the matrix $\boldsymbol{\Omega}^*$ encodes the conditional independence structure of the Gaussian graphical model, i.e. an edge is present between nodes j and k iff $\boldsymbol{X}^j \perp\!\!\!\perp \boldsymbol{X}^k | \boldsymbol{X}^{-\{j,k\}}$ which is equivalent to $\boldsymbol{\Omega}^*_{jk} = 0$. See for example[53] where this fact is shown in a more general setting. This motivates the need for estimation of $\boldsymbol{\Omega}^*$. Let $\boldsymbol{\Sigma}_n = \frac{1}{n}\sum_{i=1} \boldsymbol{X}_i^{\otimes 2}$ be the sample covariance of $\boldsymbol{X}_1, ..., \boldsymbol{X}_n$. The CLIME estimator of $\boldsymbol{\Omega}^*$ is given by:

$$\widehat{\boldsymbol{\Omega}} = \text{argmin} \|\boldsymbol{\Omega}\|_1, \text{ st } \|\boldsymbol{\Sigma}_n \boldsymbol{\Omega} - \mathbf{I}_d\|_{\max} \leq \lambda. \tag{4.5.1}$$

In this section we are interested in testing whether a precision matrix element $\boldsymbol{\Omega}^*_{1m} = 0$. In the case when $\boldsymbol{X}_i$ are coming form a Gaussian distribution this test translates to a test of whether there is an edge between nodes 1 and m in the conditional independence graph. Let $\boldsymbol{\beta}^* := \boldsymbol{\Omega}^*_{*m}$, be the m^{th}

column of $\mathbf{\Omega}^*$. Then, the CLIME reduces to

$$\widehat{\boldsymbol{\beta}} = \text{argmin} \|\boldsymbol{\beta}\|_1, \text{ st } \|\mathbf{\Sigma}_n \boldsymbol{\beta} - \mathbf{e}_m^T\|_\infty \leq \lambda.$$

Note that here $\mathbf{e}_m^T$ is a column vector. In other words, if we phrase this problem in the terminology of Section 4.3.5, we have $A_b(\boldsymbol{X}) = \boldsymbol{X}^{\otimes 2}$, and $B_b(\boldsymbol{X}) = \mathbf{e}_m^T$. According to the formulation of our test statistic, we have:

$$\widehat{S}(\boldsymbol{\beta}) = \widehat{\mathbf{v}}^T(\mathbf{\Sigma}_n \boldsymbol{\beta} - \mathbf{e}_m^T),$$

where

$$\widehat{\mathbf{v}} = \text{argmin} \|\mathbf{v}\|_1, \text{ st } \|\mathbf{v}^T \mathbf{\Sigma}_n - \mathbf{e}_1\|_\infty \leq \lambda'.$$

Remark 4.5.1. *Notice here that due to the apparent symmetry, it suffices to simply solve the CLIME optimization (4.5.1) once in order for us to perform inference, as $\widehat{\boldsymbol{\beta}} = \widehat{\mathbf{\Omega}}_{*m}$ and $\widehat{\mathbf{v}} = \widehat{\mathbf{\Omega}}_{*1}$, provided that λ and λ' can be selected to be the same.*

Similarly to the Dantzig Selector case, we assume that the coordinates of the covariates have sub-Gaussian distributions with $\sup_i \|\boldsymbol{X}^i\|_{\psi_2} \leq K_{\boldsymbol{X}}$, and furthermore that the m^{th} column of $\mathbf{\Omega}^*$ is $\boldsymbol{\beta}^*$ with $\|\boldsymbol{\beta}^*\|_0 = s$. Denote the 1^{st} column of $\mathbf{\Omega}^*$ with $\mathbf{v}^*$ and assume that $\|\mathbf{v}^*\|_0 = s_{\mathbf{v}}$. Below we provide an influence function expansion, similar to the one in the Dantzig Selector case:

Theorem 4.5.2. *Assume that the covariate distribution is sub-Gaussian, with ψ_2 norms as specified in the Dantzig selector case. Furthermore, assume that the smallest eigenvalue of the covariance matrix $\lambda_{\min}(\mathbf{\Sigma}_{\boldsymbol{X}}) > \delta > 0$ is bounded away from 0. Denote with s and $s_{\mathbf{v}}$ the sparsities of the vectors $\boldsymbol{\beta}^*$ ($\boldsymbol{\beta}_1^* = \theta = 0$, $\boldsymbol{\beta}_{-1}^* = \boldsymbol{\gamma}^*$) and $\mathbf{v}^*$ correspondingly. Under the assumption that $\max(s_{\mathbf{v}}, s)\|\mathbf{v}^*\|_1 \|\boldsymbol{\beta}^*\|_1 \frac{\log d}{\sqrt{n}} = o(1)$, and large enough tuning parameters with $\lambda \asymp \|\boldsymbol{\beta}^*\|_1 \sqrt{\frac{\log d}{n}}$ and $\lambda' \asymp \|\mathbf{v}^*\|_1 \sqrt{\frac{\log d}{n}}$, we have the following asymptotic influence function expansion of the test*

statistic:

$$n^{1/2}\widehat{S}(0,\widehat{\boldsymbol{\gamma}}) = \frac{1}{\sqrt{n}}\mathbf{v}^{*T}\left(\sum_{i=1}^{n}\boldsymbol{X}_i\boldsymbol{X}_{i,-1}^T\boldsymbol{\gamma}^* - \mathbf{e}_m^T\right) + o_p(1)$$

Remark 4.5.3. *Note here that the quantities λ and λ' are guaranteed to be $o(1)$, since $\|\mathbf{v}^*\|_1 \geq (\boldsymbol{\Sigma}_{\boldsymbol{X}}^{-1})_{11} \geq (\boldsymbol{\Sigma}_{\boldsymbol{X},11})^{-1} \geq (2K_{\boldsymbol{X}}^2)^{-1} > 0$, and similarly $\|\boldsymbol{\beta}^*\|_1 \geq (\boldsymbol{\Sigma}_{\boldsymbol{X},11})^{-1}$.*

The proof of Theorem 4.5.2 can be found in Appendix C.3.1. Next we provide a weak convergence result, whose proof is also deferred to Appendix C.3.1.

Corollary 4.5.4. *Under the same assumptions as in Theorem 4.5.2, $\frac{(s_{\mathbf{v}}s)^{3/2}}{n^{1/2}} = o(1)$ and the following assumption on the $\boldsymbol{X}$ distribution:*

$$\frac{\mathrm{Var}(\mathbf{v}^{*T}\boldsymbol{X}^{\otimes 2}\boldsymbol{\beta}^*)}{\|\boldsymbol{\beta}^*\|_2^2\|\mathbf{v}^*\|_2^2} \geq \iota_{\min} > 0, \tag{4.5.2}$$

we have

$$\frac{n^{1/2}}{\sqrt{\Delta}}\widehat{S}(0,\widehat{\boldsymbol{\gamma}}) \rightsquigarrow N(0,1), \quad \Delta = \mathrm{Var}(\mathbf{v}^{*T}\boldsymbol{X}^{\otimes 2}\boldsymbol{\beta}^*). \tag{4.5.3}$$

Note. $\boldsymbol{\beta}^* = (0, \boldsymbol{\gamma}^{*T})^T$.

Remark 4.5.5. *Here we discuss the variance assumption (4.5.2), in the case when $\boldsymbol{X} \sim N(0, \boldsymbol{\Sigma})$. By Isserlis' theorem, for any two vectors $\boldsymbol{\xi}$ and $\boldsymbol{\theta}$ we have:*

$$\begin{aligned}\mathrm{Var}(\boldsymbol{\xi}^T\boldsymbol{X}^{\otimes 2}\boldsymbol{\theta}) &= (\boldsymbol{\xi}^T\boldsymbol{\Sigma}\boldsymbol{\xi})(\boldsymbol{\theta}^T\boldsymbol{\Sigma}\boldsymbol{\theta}) + (\boldsymbol{\xi}^T\boldsymbol{\Sigma}\boldsymbol{\theta})^2 \\ &\geq \lambda_{\min}^2(\boldsymbol{\Sigma})\|\boldsymbol{\xi}\|_2^2\|\boldsymbol{\theta}\|_2^2,\end{aligned}$$

which clearly implies (4.5.2). Furthermore, we note that in our setting assumption (4.5.2) is equivalent to $\mathrm{Var}(\mathbf{v}^{*T}\boldsymbol{X}^{\otimes 2}\boldsymbol{\beta}^*) \geq V_{\min} > 0$, as we know from Remark 4.4.4 $\|\mathbf{v}^*\|_2, \|\boldsymbol{\beta}^*\|_2 \leq \delta^{-1}$ and $\|\mathbf{v}^*\|_2 \geq |\mathbf{v}_{11}^*| \geq (2K_{\boldsymbol{X}}^2)^{-1}$ and similarly $\|\boldsymbol{\beta}^*\|_2 \geq (2K_{\boldsymbol{X}}^2)^{-1}$.

Obviously, there are two intuitive plugin type of estimators for the variance Δ, as defined in (4.5.3) — $\widehat{\Delta}_1 := \frac{1}{n}\sum_{i=1}^n(\widehat{\mathbf{v}}^T(\boldsymbol{X}_i^{\otimes 2} - \boldsymbol{\Sigma}_n)\widehat{\boldsymbol{\beta}})^2$ and $\widehat{\Delta}_2 = \frac{1}{n}\sum_{i=1}^n(\widehat{\mathbf{v}}^T\boldsymbol{X}_i^{\otimes 2}\widehat{\boldsymbol{\beta}} - \widehat{\mathbf{v}}^T\mathbf{e}_m^T)^2$, where we recall that e_m^T is a row unit vector. Next we show both estimators are consistent under certain conditions, starting with the former one.

Proposition 4.5.6. *Under the assumptions from Theorem 4.5.2, and the following additional assumptions:*

$$\max(s_{\mathbf{v}}\|\mathbf{v}^*\|_1, s\|\boldsymbol{\beta}^*\|_1)\|\mathbf{v}^*\|_1\|\boldsymbol{\beta}^*\|_1\sqrt{\frac{\log d}{n}}\log(nd) = o(1),$$
$$\mathrm{Var}\left[\left(\mathbf{v}^{*T}\boldsymbol{X}^{\otimes 2}\boldsymbol{\beta}^*\right)^2\right] = o(n),$$

we have that the plugin estimator $\widehat{\Delta}_1$ is consistent for Δ, as defined in (4.5.3).

Remark 4.5.7. *The alternative estimator $\widehat{\Delta}_2$ is also consistent, under same assumptions as in Proposition 4.5.6.*

Remark 4.5.8. *The variance condition in Proposition 4.5.6 is trivially satisfied if the variance is finite. Note that by Cauchy-Schwartz this is true if:*

$$\mathbb{E}\left(\mathbf{v}^{*T}\boldsymbol{X}\right)^8 < \infty, \quad \mathbb{E}\left(\boldsymbol{\beta}^{*T}\boldsymbol{X}\right)^8 < \infty,$$

which is in turn trivially satisfied if $\mathbf{v}^{*T}\boldsymbol{X}$ and $\boldsymbol{\beta}^{*T}\boldsymbol{X}$ are sub-Gaussian, and is obvious in the case when $\boldsymbol{X}$ is multivariate normal.

The proofs of Proposition 4.5.6 and Remark 4.5.7 are provided in Appendix C.3.1. Proposition 4.5.6 and Remark 4.5.7 suggest estimates $\widehat{\Delta}_1$ and $\widehat{\Delta}_2$ satisfy the condition in Proposition 4.3.6, and hence the two statistics based on them will provide results with correct size under the null distribution. Hence we can state the following:

Theorem 4.5.9. *Assume all assumptions in Corollary 4.5.4, and construct estimates of* $\Delta - \widehat{\Delta}_1$ *and* $\widehat{\Delta}_2$ *based on Propositions 4.5.6 and Remark 4.5.7 under their corresponding conditions. Then the statistics* $\widehat{U}_n^i = \frac{n^{1/2}}{\sqrt{\widehat{\Delta}_i}} \widehat{S}(0, \widehat{\boldsymbol{\gamma}})$ *satisfy:*

$$\lim_{n\to\infty} |\mathbb{P}^*(\widehat{U}_n^i \leq t) - \Phi(t))| = 0, \;\; i = 1, 2.$$

Proof of Theorem 4.5.9. The proof of this theorem follows from the previous statements in this Section showing that the conditions of Proposition 4.3.6 are satisfied. □

Next we proceed to formulate a uniform weak convergence result. To this end for a fixed $M > \delta > 0$, define the following parameter space of covariance matrices:

$$\mathcal{S}_0(L, s) = \{\boldsymbol{\Sigma} : \boldsymbol{\Sigma} = \boldsymbol{\Sigma}^T, 0 < \delta \leq \boldsymbol{\Sigma}, \|\boldsymbol{\Sigma}\|_{\max} \leq M, \boldsymbol{\Sigma}_{1m} = 0, \|\boldsymbol{\Sigma}^{-1}\|_1 \leq L, \max_i \|\boldsymbol{\Sigma}_{*i}^{-1}\|_0 \leq s\}.$$

Remark 4.5.10. *Observe that for a matrix* $\boldsymbol{\Sigma} \in \mathcal{S}_0(L, s)$ *we have* $\max_i \|\boldsymbol{\Sigma}_{*i}^{-1}\|_2 \leq \delta^{-1}$*, since:*

$$(\boldsymbol{\Sigma}^{-1})_{ii} = \boldsymbol{\Sigma}_{*i}^{-1T} \boldsymbol{\Sigma} \boldsymbol{\Sigma}_{*i}^{-1} \geq \|\boldsymbol{\Sigma}_{*i}^{-1}\|_2^2 \delta \geq (\boldsymbol{\Sigma}^{-1})_{ii}^2 \delta.$$

If $(\boldsymbol{\Sigma}^{-1})_{ii}^2 = 0$ *in the above inequality, trivially we have* $\|\boldsymbol{\Sigma}_{*i}^{-1}\|_2 = 0$*. Otherwise, it follows* $\|\boldsymbol{\Sigma}_{*i}^{-1}\|_2 \leq \delta^{-1}$*. Moreover* $(\boldsymbol{\Sigma}_X^{-1})_{ii} \geq M^{-1}$ *as we argued in Remark 4.4.4, and hence by Cauchy-Schwartz* $L \leq \sqrt{s}\delta^{-1}$*. Also obviously* $M \leq Ls$*.*

We have the following theorem in terms of uniform convergence:

Theorem 4.5.11. *Let* $\boldsymbol{X}$ *belong to a sub-Gaussian distribution with* $\sup_i \|\boldsymbol{X}^i\|_{\psi_2} \leq K_{\boldsymbol{X}}$*, for some* $(M/2)^{1/2} \leq K_{\boldsymbol{X}} < \infty$ *and* $\mathrm{Cov}(\boldsymbol{X}) = \boldsymbol{\Sigma}_{\boldsymbol{X}} \in \mathcal{S}_0(L, s)$*. Let* $\boldsymbol{\Omega} = (\boldsymbol{\Sigma}_{\boldsymbol{X}})^{-1}$ *and denote with*

$\boldsymbol{\beta} = \boldsymbol{\Omega}_{*m}$, and $\mathbf{v} = \boldsymbol{\Omega}_{*1}$. We assume the following moment conditions on the $\boldsymbol{X}$ distribution:

$$\mathrm{Var}(\mathbf{v}^T \boldsymbol{X}^{\otimes 2} \boldsymbol{\beta}) \geq V_{\min} > 0, \quad \mathrm{Var}((\mathbf{v}^T \boldsymbol{X}^{\otimes 2} \boldsymbol{\beta})^2) \leq V_{\max} < \infty. \tag{4.5.4}$$

Then under the following conditions:

$$sL^2 \frac{\log(d)}{\sqrt{n}} = o(1), \quad sL^3 \sqrt{\frac{\log(d)}{n}} \log(nd) = o(1), \quad \frac{s^3}{\sqrt{n}} = o(1), \tag{4.5.5}$$

we have:

$$\lim_{n \to \infty} \sup_{\boldsymbol{\Sigma}_{\boldsymbol{X}} \in \mathcal{S}_0(L,s)} \sup_t |\mathbb{P}_{\boldsymbol{\beta}}(\widehat{U}_n^i \leq t) - \Phi(t)| = 0, i = 1, 2.$$

Remark 4.5.12. *Note that the sub-Gaussian assumption on $\boldsymbol{X}$ implies that $\|\boldsymbol{\Sigma}\|_{\max} \leq 2K_{\boldsymbol{X}}^2$, hence the requirement on $K_{\boldsymbol{X}}$ with respect to M.*

Next we formulate a result on the local power. Similarly to above we construct the following parameter space:

$$\mathcal{S}_1(K, \phi, L, s) = \{\boldsymbol{\Sigma} : \boldsymbol{\Sigma} = \boldsymbol{\Sigma}^T, 0 < \delta \leq \boldsymbol{\Sigma}, \|\boldsymbol{\Sigma}\|_{\max} \leq M, \boldsymbol{\Sigma}_{1m} = Kn^{-\phi}, \|\boldsymbol{\Sigma}^{-1}\|_1 \leq L, \max_i \|\boldsymbol{\Sigma}_{*i}^{-1}\|_0 \leq s\}.$$

We then have the following:

Theorem 4.5.13. *Let X belong to a sub-Gaussian distribution with $\sup_i \|\boldsymbol{X}^i\|_{\psi_2} \leq K_{\boldsymbol{X}}$, for some $(M/2)^{1/2} \leq K_{\boldsymbol{X}} < \infty$ and $\mathrm{Cov}(\boldsymbol{X}) = \boldsymbol{\Sigma}_{\boldsymbol{X}} \in \mathcal{S}_1(K, \phi, L, s)$. Furthermore assume that the moment conditions (4.5.4). Under the assumption $\max(1, Ms) KLn^{-\phi} \sqrt{\log d} = o(1)$, and*

assumptions (4.5.5), we have that for $i = 1, 2$:

$$\lim_{n\to\infty} \sup_{\mathbf{\Sigma}_{\boldsymbol{X}} \in \mathcal{S}_1(K,\phi,L,s)} \sup_t \left| \mathbb{P}_{\boldsymbol{\beta}}(\widehat{U}_n^i \leq t) - \Phi(t) \right| = 0, \ \textit{if}\, \phi > 1/2,$$
$$\lim_{n\to\infty} \sup_{\mathbf{\Sigma}_{\boldsymbol{X}} \in \mathcal{S}_1(K,\phi,L,s)} \sup_t \left| \mathbb{P}_{\boldsymbol{\beta}}(\widehat{U}_n^i \leq t) - \Phi\left(t + \frac{K}{\sqrt{\Delta}}\right) \right| = 0, \ \textit{if}\, \phi = 1/2,$$

and for a fixed $t \in \mathbb{R}$ and $K \neq 0$ we have:

$$\lim_{n\to\infty} \sup_{\mathbf{\Sigma}_{\boldsymbol{X}} \in \mathcal{S}_1(K,\phi,L,s)} \mathbb{P}_{\boldsymbol{\beta}}(|\widehat{U}_n^i| \leq t) = 0, \ \textit{if}\, \phi < 1/2.$$

The proof of this theorem is left to appendix C.3.1.

One-step Estimator and Confidence Intervals

Following Section 4.3.4, we can define the one-step estimator as (assuming WLOG $j = 1$):

$$\widetilde{\theta} = \widehat{\theta} - \frac{\widehat{\mathbf{v}}^T(\mathbf{\Sigma}_n \widehat{\boldsymbol{\beta}} - \mathbf{e}_m^T)}{\widehat{\mathbf{v}}^T \sum_{i=1}^n \boldsymbol{X}_i \boldsymbol{X}_{i,1}/n}. \tag{4.5.6}$$

Next we show the following:

Corollary 4.5.14. *Under the assumptions of Corollary 4.5.4, we have:*

$$\frac{1}{\sqrt{\Delta}} n^{1/2}(\widetilde{\theta} - \theta^*) \rightsquigarrow N(0, 1)$$

where Δ is defined as in (4.5.3).

Proof. This simply follows from Proposition 4.3.26, upon noting that $\lambda' \asymp \|\mathbf{v}^*\|_1 \sqrt{\frac{\log d}{n}} = o(1)$. □

Remark 4.5.15. *As it becomes evident from Corollaries 4.5.4 and 4.5.14 the asymptotic distributions of the suggested test and the one-step estimator test (based on $\widetilde{\theta}$) are asymptotically equivalent.*

Remark 4.5.16. *Clearly we can use the plugin estimator suggested in Proposition 4.5.6, to construct confidence intervals, or test the parameter using the one-step estimator approach suggested above.*

4.5.2 Transelliptical Graphical Models with CLIME

In this subsection we consider a related framework to the CLIME example above, namely we consider the transelliptical graphical models (TGM), proposed by Liu et al.[52]. We recall several definitions before we proceed.

Definition 4.5.17 (elliptical distribution Fang et al.[21])**.** *Let $\boldsymbol{\mu} \in \mathbb{R}^d$ and $\boldsymbol{\Sigma} \in \mathbb{R}^{d\times d}$. We say that the d-dimesnional vector $\boldsymbol{X}$ has an elliptical distribution, and we denote it with $\boldsymbol{X} \sim EC_d(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \xi)$ if $\boldsymbol{X} \overset{d}{=} \boldsymbol{\mu} + \xi A\boldsymbol{U}$, where $\boldsymbol{U}$ is a random vector uniformly distributed on the unit sphere in $\mathbb{R}^q$, $\xi \geq 0$ is a scalar random variable independent of $\boldsymbol{U}$, $A \in \mathbb{R}^{d\times q}$ is a deterministic matrix such that $AA^T = \boldsymbol{\Sigma}$.*

Definition 4.5.18 (transelliptical distribution Liu et al.[52])**.** *We call the continuous random vector $\boldsymbol{X} = (\boldsymbol{X}^1, \ldots, \boldsymbol{X}^d)^T$ transelliptically distributed, and we denote it with $\boldsymbol{X} \sim TE_d(\boldsymbol{\Sigma}, \xi; f_1, \ldots, f_d)$, if there exists a set of monotone univariate functions $f_1, \ldots, f_d$ and a non-negative random variable ξ, with $\mathbb{P}(\xi = 0) = 0$, such that:*

$$(f_1(\boldsymbol{X}^1), \ldots, f_d(\boldsymbol{X}^d))^T \sim EC_d(0, \boldsymbol{\Sigma}, \xi),$$

where $\boldsymbol{\Sigma}$ is symmetric with $diag(\boldsymbol{\Sigma}) = 1$ and $\boldsymbol{\Sigma} > 0$ in a positive-definite sense. Here $\boldsymbol{\Sigma}$ is called the "latent generalized correlation matrix".

It is worth mentioning that the family of transelliptical distributions (TD) is a broad family of distributions, subsuming the family of nonparanormal distributions e.g., the definition of which can be found in Liu et al.[51].

The graphical structure in TGMs can then be defined through the notion of the "latent generalized concentration matrix" — $\boldsymbol{\Omega} = \boldsymbol{\Sigma}^{-1}$, i.e. an edge is present between two variables $\boldsymbol{X}^j, \boldsymbol{X}^k$ iff $\boldsymbol{\Omega}_{jk} \neq 0$. Such a construction extends, classical results from the Gaussian graphical models (for more details see Lemma 3.2 and Lemma 3.3 in Liu et al.[52]).

To construct an estimate of $\boldsymbol{\Omega}$, Liu et al.[52] suggest estimating the correlation matrix $\boldsymbol{\Sigma}$ first. This can be done by using a non-parametric estimate of the correlation such as Kendall's tau statistic, and transforming it back, to obtain an estimate of $\boldsymbol{\Sigma}$.

Assume again that $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ are iid copies of $\boldsymbol{X}$. Kendall's tau statistic is a matrix whose jk^{th} element is:

$$\widehat{\tau}_{jk} = \frac{2}{n(n-1)} \sum_{1 \leq i < i' \leq n} \operatorname{sign}\left((\boldsymbol{X}_i^j - \boldsymbol{X}_{i'}^j)(\boldsymbol{X}_i^k - \boldsymbol{X}_{i'}^k)\right).$$

To get an estimate of the correlation matrix we then transform the $\widehat{\tau}_{jk}$. Note that it is clear from the definition of $\widehat{\tau}_{jk}$, that it is an unbiased estimator of:

$$\tau_{jk} = \mathbb{P}((\boldsymbol{Y}^j - \boldsymbol{Y}'^j)(\boldsymbol{Y}^k - \boldsymbol{Y}'^k) > 0) - \mathbb{P}((\boldsymbol{Y}^j - \boldsymbol{Y}'^j)(\boldsymbol{Y}^k - \boldsymbol{Y}'^k) < 0),$$

where $\boldsymbol{Y}, \boldsymbol{Y}' \sim \boldsymbol{X}$ are iid random variables. Define:

$$\widehat{\boldsymbol{S}}_{jk}^{\tau} = \begin{cases} \sin\left(\frac{\pi}{2}\widehat{\tau}_{jk}\right), & j \neq k; \\ 1, & j = k. \end{cases}$$

It can be seen that $\widehat{\boldsymbol{S}}_{jk}^{\tau}$ consistently estimates $\boldsymbol{\Sigma}$ (see Theorem 4.5.19 below). Let $\boldsymbol{\Omega}^* = \boldsymbol{\Sigma}^{-1}$. The

TGM estimator with CLIME is given by:

$$\widehat{\boldsymbol{\Omega}} = \text{argmin} \, \|\boldsymbol{\Omega}\|_1, \;\; \text{st} \;\; \|\widehat{\boldsymbol{S}}^\tau \boldsymbol{\Omega} - \mathbf{I}_d\|_{\max} \leq \lambda.$$

To test whether the element of the matrix $\boldsymbol{\Omega}^*_{1m} = 0$, we can apply the same approach as in CLIME. Denote with $\boldsymbol{\beta} = \boldsymbol{\Omega}_{*m}$. Then, the CLIME with TGM reduces to

$$\widehat{\boldsymbol{\beta}} = \text{argmin} \, \|\boldsymbol{\beta}\|_1, \;\; \text{st} \;\; \|\widehat{\boldsymbol{S}}^\tau \boldsymbol{\beta} - \mathbf{e}_m^T\|_\infty \leq \lambda.$$

According to our formulation of the test statistic we have:

$$\widehat{S}(\boldsymbol{\beta}) = \widehat{\mathbf{v}}^T (\widehat{\boldsymbol{S}}^\tau \boldsymbol{\beta} - \mathbf{e}_m^T),$$

where

$$\widehat{\mathbf{v}} = \text{argmin} \, \|\mathbf{v}\|_1, \;\; \text{st} \;\; \|\mathbf{v}^T \widehat{\boldsymbol{S}}^\tau - \mathbf{e}_1\|_\infty \leq \lambda'.$$

We note that the structure in the TGM with CLIME is slightly different than the one we suggested in Section 4.3.5, due to the U-statistic structure of $\boldsymbol{S}^\tau$ as compared to estimating equations before which had iid structure. Nevertheless, we can still show that the asymptotic theory goes through in this case.

We will show the normality of the our test statistic in this setting below. Note that we can no longer use the lemmas from the CLIME case, as the estimator of $\boldsymbol{\Sigma}$ is constructed in a completely different manner. Furthermore, the vector $\boldsymbol{X}$, coming from a nonparanormal family need not be sub-Gaussian. Fortunately enough, Liu et al.[51] provide a theorem stated below, with the help of which we can show the normality:

Theorem 4.5.19 (Liu 2012). *For any $n > 1$ with probability at least $1 - 1/d$, we have*

$$\|\widehat{\boldsymbol{S}}^{\tau} - \boldsymbol{\Sigma}\|_{\max} \leq 2.45\pi\sqrt{\frac{\log d}{n}}. \tag{4.5.7}$$

While this theorem is defined within the framework of nonparanormal models, the proof doesn't utilize the fact that the family is nonparanormal, and thus extends to the transelliptical case. As we can see from the theorem, the rate of Kendall's tau estimate (4.5.7), is no different than the one using the sample covariance matrix, provided in Lemma C.2.2.

We are now in position to formulate the influence function expansion of the test statistic in the TGM with CLIME:

Theorem 4.5.20. *Assume the covariate distribution follows a nonparanormal model with a function f and correlation matrix $\boldsymbol{\Sigma}$. Furthermore, assume that the smallest eigenvalue of the correlation matrix satisfies $\lambda_{\min}(\boldsymbol{\Sigma}) > \delta > 0$ is bounded away from o, and diag$(\boldsymbol{\Sigma}) = 1$. Let $\|\boldsymbol{\beta}^*\|_0 = s$ and $\|\mathbf{v}^*\|_0 = s_{\mathbf{v}}$. Under the assumption that $\max(s_{\mathbf{v}}, s)\|\mathbf{v}^*\|_1\|\boldsymbol{\beta}^*\|_1\frac{\log d}{\sqrt{n}} = o(1)$, and large enough tuning parameters with $\lambda \asymp \|\boldsymbol{\beta}^*\|_1\sqrt{\frac{\log d}{n}}$ and $\lambda' \asymp \|\mathbf{v}^*\|_1\sqrt{\frac{\log d}{n}}$, we have the following asymptotic influence function expansion of the test statistic:*

$$n^{1/2}\widehat{S}(0, \widehat{\boldsymbol{\gamma}}) = n^{1/2}\mathbf{v}^{*T}\left(\widehat{\boldsymbol{S}}^{\tau}\boldsymbol{\beta}^* - \mathbf{e}_m^T\right) + o_p(1)$$

Note. *$\boldsymbol{\beta}^* = (0, \boldsymbol{\gamma}^{*T})^T$, with the o being on the 1^{st} place.*

Proof. The proof is identical to the one for the CLIME, up to usages of Lemma C.3.6 instead of Lemma C.2.5 and thus we omit it. □

Next we formulate a theorem akin to Corollary 4.5.4. The proof of this theorem is technical and we defer it to Appendix C.3.2. The proof relies on the Hájek projection approach and Hoeffding's

U-statistic inequality, and is different in spirit to Corollary 4.3.5, as the U-statistic structure no longer allows the simple iid decomposition, which we had before.

Theorem 4.5.21. *Assume the same assumptions as in Theorem 4.5.20, and furthermore, assume that* $\frac{(s_{\mathbf{v}}s)^{3/2}}{n^{1/2}} = o(1)$, $\frac{\sqrt{s_{\mathbf{v}}s}\log d}{n^{1/2}} = o(1)$ *and that:*

$$\mathrm{Var}(\mathbf{v}^{*T}\Theta\boldsymbol{\beta}^*) \geq \iota_{\min}\|\mathbf{v}^*\|_2^2\|\boldsymbol{\beta}^*\|_2^2, \quad \iota_{\min} > 0$$

where Θ *is a* $d \times d$ *random matrix with entries* $\Theta_{jk} = \pi\cos\left(\frac{\pi}{2}\tau_{jk}\right)\tau_{jk}^{\boldsymbol{Y}}$*, where:*

$$\tau_{jk}^{\boldsymbol{Y}} = [\mathbb{P}((\boldsymbol{Y}^j - \boldsymbol{Y}'^j)(\boldsymbol{Y}^k - \boldsymbol{Y}'^k) > 0|\boldsymbol{Y}) - \mathbb{P}((\boldsymbol{Y}^j - \boldsymbol{Y}'^j)(\boldsymbol{Y}^k - \boldsymbol{Y}'^k) < 0|\boldsymbol{Y}) - \tau_{jk}],$$

with $\boldsymbol{Y}, \boldsymbol{Y}'$ *are iid copies of* $\sim \boldsymbol{X}$ *(and all* $\tau_{jk}^{\boldsymbol{Y}}$ *being a random variable depending on* $\boldsymbol{Y}$*).*

Then we have that:

$$\frac{n^{1/2}}{\sqrt{\Delta}}\widehat{S}(0,\widehat{\boldsymbol{\gamma}}) \rightsquigarrow N(0,1), \text{ where } \Delta = \mathrm{Var}(\mathbf{v}^{*T}\Theta\boldsymbol{\beta}^*). \tag{4.5.8}$$

Remark 4.5.22. *As in the CLIME case, we can show that the condition* $\mathrm{Var}(\mathbf{v}^{*T}\Theta\boldsymbol{\beta}^*) \geq \iota_{\min}\|\mathbf{v}^*\|_2^2\|\boldsymbol{\beta}^*\|_2^2$ *is equivalent to* $\mathrm{Var}(\mathbf{v}^{*T}\Theta\boldsymbol{\beta}^*) \geq V_{\min}$.

Next we turn our attention to consistent estimation of Δ as defined in (4.5.8). To this end define the following matrices — $\widehat{\Theta}^i$:

$$\widehat{\tau}_{jk}^i = \frac{1}{n-1}\sum_{i'\neq i}\mathrm{sign}\left((\boldsymbol{X}_i^j - \boldsymbol{X}_{i'}^j)(\boldsymbol{X}_i^k - \boldsymbol{X}_{i'}^k)\right) - \widehat{\tau}_{jk},$$
$$\widehat{\Theta}_{jk}^i = \pi\cos\left(\frac{\pi}{2}\widehat{\tau}_{jk}\right)\widehat{\tau}_{jk}^i.$$

Note that $\widehat{\Theta}_{jk}^i$ is symmetric by definition. Define the estimator $\widehat{\Delta} = \frac{1}{n}\sum_{i=1}^n(\widehat{\mathbf{v}}^T\widehat{\Theta}^i\widehat{\boldsymbol{\beta}})^2$. We have

the following:

Proposition 4.5.23. *Under the same assumptions as in Theorem 4.5.21, and the additional assumptions:*

$$\|\mathbf{v}^*\|_1^2\|\boldsymbol{\beta}^*\|_1^2\sqrt{\frac{\log(nd)}{n}} = o(1),$$
$$\|\mathbf{v}^*\|_1^2\|\boldsymbol{\beta}^*\|_1^2 \max(s_{\mathbf{v}}, s)\sqrt{\frac{\log d}{n}} = o(1),$$
$$\mathrm{Var}((\mathbf{v}^*\Theta\boldsymbol{\beta}^*)^2) = o(n),$$

we have that $\widehat{\Delta} \to_p \Delta$.

Remark 4.5.24. *As we saw before* $\|\mathbf{v}^*\|_2, \|\boldsymbol{\beta}^*\|_2 \leq \delta^{-1}$. *Since the elements of* Θ *are bounded by* 2π *we have* $|\mathbf{v}^{*T}\Theta\boldsymbol{\beta}^*| \leq \delta^{-2}\sqrt{s_{\mathbf{v}}s}2\pi$. *Hence a sufficient condition for the variance condition is* $\frac{(s_{\mathbf{v}}s)^2}{n} = o(1)$. *Note that this has already been assumed.*

Let $\widehat{U}_n = \frac{n^{1/2}}{\sqrt{\widehat{\Delta}}}\widehat{S}(0, \widehat{\boldsymbol{\gamma}})$. Combining the results from Proposition 4.5.23 and Theorem 4.5.9, we get the following:

Theorem 4.5.25. *Assume all assumptions in Proposition 4.5.23 and Theorem 4.5.9. Then the statistic* $\widehat{U}_n = \frac{n^{1/2}}{\sqrt{\widehat{\Delta}}}\widehat{S}(0, \widehat{\boldsymbol{\gamma}})$ *satisfies:*

$$\lim_{n\to\infty} |\mathbb{P}^*(\widehat{U}_n \leq t) - \Phi(t))| = 0.$$

Furthermore, similarly to the CLIME section above, we can show the following two results uniform weak convergence and local power. Before that we define two classes of correlation matrices, in

analogy to the CLIME case:

$$\widetilde{\mathcal{S}}_0(L,s) = \{\mathbf{\Sigma} : \mathbf{\Sigma} = \mathbf{\Sigma}^T, 0 < \delta \leq \mathbf{\Sigma}, \text{diag}(\mathbf{\Sigma}) = 1, \mathbf{\Sigma}_{1m} = 0, \|\mathbf{\Sigma}^{-1}\|_1 \leq L, \max_i \|\mathbf{\Sigma}_{*i}^{-1}\|_0 \leq s\},$$

$$\widetilde{\mathcal{S}}_1(K,\phi,L,s) = \{\mathbf{\Sigma} : \mathbf{\Sigma} = \mathbf{\Sigma}^T, 0 < \delta \leq \mathbf{\Sigma}, \text{diag}(\mathbf{\Sigma}) = 1, \mathbf{\Sigma}_{1m} = Kn^{-\phi}, \|\mathbf{\Sigma}^{-1}\|_1 \leq L, \max_i \|\mathbf{\Sigma}_{*i}^{-1}\|_0 \leq s\}.$$

Theorem 4.5.26. *Let $X \sim TE_d(\boldsymbol{\mu}, \mathbf{\Sigma}, \xi)$ with $\mathbf{\Sigma} \in \widetilde{\mathcal{S}}_0(L,s)$. where Θ is defined in Theorem 4.5.21. Assume furthermore that X satisfies the following moment conditions:*

$$\text{Var}(\mathbf{v}^T\Theta\boldsymbol{\beta}) \geq V_{\min} > 0, \quad \text{Var}((\mathbf{v}^T\Theta\boldsymbol{\beta})^2) \leq V_{\max} < \infty, \tag{4.5.9}$$

and

$$s^3 n^{-1/2} = o(1), \quad s\log(d)n^{-1/2} = o(1), \quad sL^2\log(d)n^{-1/2} = o(1), \tag{4.5.10}$$

$$L^4\sqrt{\frac{\log(nd)}{n}} = o(1), \quad L^4 s\sqrt{\frac{\log d}{n}} = o(1), \tag{4.5.11}$$

we have:

$$\lim_{n\to\infty} \sup_{\mathbf{\Sigma}\in\widetilde{\mathcal{S}}_0(L,s)} \sup_t |\mathbb{P}_{\boldsymbol{\beta}}(\widehat{U}_n \leq t) - \Phi(t)| = 0.$$

The result on local power is formulated below.

Theorem 4.5.27. *Assume the same assumptions as in Theorem 4.5.26 and in addition $sKLn^{-\phi}\sqrt{\log d} = o(1)$. Then we have:*

$$\lim_{n\to\infty} \sup_{\mathbf{\Sigma}_{\boldsymbol{X}}\in\widetilde{\mathcal{S}}_1(K,\phi,L,s)} \sup_t \left|\mathbb{P}_{\boldsymbol{\beta}}(\widehat{U}_n \leq t) - \Phi(t)\right| = 0, \ \textit{if}\, \phi > 1/2,$$

$$\lim_{n\to\infty} \sup_{\mathbf{\Sigma}_{\boldsymbol{X}}\in\widetilde{\mathcal{S}}_1(K,\phi,L,s)} \sup_t \left|\mathbb{P}_{\boldsymbol{\beta}}(\widehat{U}_n \leq t) - \Phi\left(t + \frac{K}{\sqrt{\Delta}}\right)\right| = 0, \ \textit{if}\, \phi = 1/2,$$

and for a fixed $t \in \mathbb{R}$ and $K \neq 0$ we have:

$$\lim_{n\to\infty} \sup_{\mathbf{\Sigma}_{\boldsymbol{X}} \in \widetilde{\mathcal{S}}_1(K,\phi,L,s)} \mathbb{P}_{\boldsymbol{\beta}}(|\widehat{U}_n| \leq t) = 0, \ \textit{if}\, \phi < 1/2.$$

One-step Estimator and Confidence Intervals

Analogously to the CLIME case (4.5.6), we have the following one step estimator on the TGM with CLIME case:

$$\widetilde{\theta} = \widehat{\theta} - \frac{\widehat{\mathbf{v}}^T(\widehat{\boldsymbol{S}}^\tau \widehat{\boldsymbol{\beta}} - \mathbf{e}_m^T)}{\widehat{\mathbf{v}}^T \widehat{\boldsymbol{S}}_{*1}^\tau}. \tag{4.5.12}$$

where $\widehat{\boldsymbol{S}}_{*1}^\tau$ is the first column (the one corresponding to the position of θ in $\boldsymbol{\beta}^*$) of $\widehat{\boldsymbol{S}}^\tau$. We next show an equivalent result to Corollary 4.5.14 in this case:

Corollary 4.5.28. *We have that:*

$$\frac{n^{1/2}}{\sqrt{\Delta}}(\widetilde{\theta} - \theta^*) \rightsquigarrow N(0,1), \tag{4.5.13}$$

where Δ is defined as in (4.5.8).

Proof. The proof is identical to the proof of Corollary 4.5.14, so we omit it. □

Remark 4.5.29. *Under the assumptions of Theorem 4.5.25 one can use $\widehat{\Delta}$ in the place of Δ to construct confidence interval in practice, and the weak convergence described in Corollary 4.5.28 still holds.*

4.6 Sparse LDA with the LDP algorithm

Another example we consider in this section, is the direct estimation for sparse LDA suggested in Cai and Liu [12]. We briefly review the problem setup below. Let $\boldsymbol{X}$ and $\boldsymbol{Y}$ are d-dimensional random vectors, coming from the same distribution centered at different means — $\boldsymbol{\mu}_1$ and $\boldsymbol{\mu}_2$ correspondingly, but sharing the same covariance matrix — $\boldsymbol{\Sigma}$. We are interested in classifying observations in population 1 or population 2. This setup has been studied extensively in the low dimensional situation. It is well known (e.g. see Mardia et al. [57] Theorem 11.2.1) that, in the case when the distribution is a multivariate normal distribution and $\boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \boldsymbol{\Sigma}$ are known, and we are drawing a new observation with equal prior probability from population 1 or 2, then the Bayes classification rule for a new observation $\boldsymbol{Z}$, takes the form:

$$\psi(\boldsymbol{Z}) = I((\boldsymbol{Z} - \boldsymbol{\mu})^T \boldsymbol{\Omega}\boldsymbol{\delta} > 0),$$

where $\boldsymbol{\mu} = (\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2)/2$, $\boldsymbol{\delta} = (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)$ and $\boldsymbol{\Omega} = \boldsymbol{\Sigma}^{-1}$. The classification rule ψ classifies the observation $\boldsymbol{Z}$ in population 1 iff $\psi(\boldsymbol{Z}) = 1$.

Clearly in practice one would never expect to know $\boldsymbol{\mu}_1, \boldsymbol{\mu}_2$ or $\boldsymbol{\Omega}$, and this renders the need for estimates of these quantities in order for the classification rule to be useful. Let us observe n_1 and n_2 samples from population 1 and population 2 correspondingly — $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_{n_1}$ and $\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_{n_2}$. Define the sample means $\bar{\boldsymbol{X}} = \frac{1}{n_1}\sum_{i=1}^{n_1} \boldsymbol{X}_i$ and $\bar{\boldsymbol{Y}} = \frac{1}{n_2}\sum_{i=1}^{n_2} \boldsymbol{Y}_i$, and the sample covariances $\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{X}} = \frac{1}{n_1}\sum_{i=1}^{n_1}(\boldsymbol{X}_i - \bar{\boldsymbol{X}})^{\otimes 2}$ and $\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{Y}} = \frac{1}{n_1}\sum_{i=1}^{n_2}(\boldsymbol{Y}_i - \bar{\boldsymbol{Y}})^{\otimes 2}$. Furthermore let $\widehat{\boldsymbol{\Sigma}}_n = \frac{n_1}{n}\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{X}} + \frac{n_2}{n}\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{Y}}$.

In the high-dimensional setting with $d \gg n$, estimates of $\boldsymbol{\Omega}$ can be unstable, given the fact that the sample covariance is not invertible. Noting that the classification rule solely depends on $\boldsymbol{\beta}^* = \boldsymbol{\Omega}\boldsymbol{\delta}$, Cai and Liu [12] suggest estimating the product of the two directly, rather than having to

estimate both of them separately. Their estimated classification rule can be summarized along the following lines:

$$\widehat{\psi}(\boldsymbol{Z}) = I((\boldsymbol{Z} - (\bar{\boldsymbol{X}} - \bar{\boldsymbol{Y}})/2)^T \widehat{\boldsymbol{\beta}} > 0), \quad \text{where}$$
$$\widehat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta} \in \mathbb{R}^d}{\operatorname{argmin}} \{ \|\boldsymbol{\beta}\|_1 : \|\widehat{\boldsymbol{\Sigma}}_n \boldsymbol{\beta} - (\bar{\boldsymbol{X}} - \bar{\boldsymbol{Y}})\|_\infty \leq \lambda \}. \tag{4.6.1}$$

In their paper Cai and Liu [12], study the properties of the classification rule $\widehat{\psi}(Z)$. Below we are interested in testing whether a certain entry of the parameter $\boldsymbol{\beta}^*$ is 0.

Define $\mathbf{v}$ as the solution:

$$\widehat{\mathbf{v}} = \operatorname{argmin} \|\mathbf{v}\|_1, \ \text{ st } \ \|\mathbf{v}^T \widehat{\boldsymbol{\Sigma}}_n - \mathbf{e}\|_\infty \leq \lambda'.$$

with $\mathbf{e}$ being a unit row vector with 1 at the position corresponding to the entry in $\boldsymbol{\beta}$ we are testing.

Define the projected test statistic in the following manner:

$$\widehat{S}(\boldsymbol{\beta}) = \widehat{\mathbf{v}}^T (\widehat{\boldsymbol{\Sigma}}_n \boldsymbol{\beta} - (\bar{\boldsymbol{X}} - \bar{\boldsymbol{Y}})).$$

Without loss of generality assume that we are interested in testing whether $\boldsymbol{\beta}^{*1} = \theta = 0$. Construct the test statistic $\widehat{S}(0, \widehat{\boldsymbol{\gamma}})$, where $\widehat{\boldsymbol{\beta}} = (\widehat{\theta}, \widehat{\boldsymbol{\gamma}}^T)^T$. We will assume that:

$$\boldsymbol{X} = \boldsymbol{\mu}_1 + \boldsymbol{U},$$
$$\boldsymbol{Y} = \boldsymbol{\mu}_2 + \boldsymbol{U},$$

where $\boldsymbol{U} = (\boldsymbol{U}^1, \ldots, \boldsymbol{U}^d)^T$ is a d-dimensional random vector, coming from a zero centered sub-Gaussian distribution as defined in the Dantzig selector section, with $\sup_i \|\boldsymbol{U}^i\|_{\psi_2} = K_{\boldsymbol{U}}$, with

covariance matrix $\mathbf{\Sigma}$. We can then represent the data of the two populations as $\boldsymbol{X}_i = \boldsymbol{\mu}_1 + \boldsymbol{U}_i, i = 1, \ldots, n_1$ and $\boldsymbol{Y}_i = \boldsymbol{\mu}_2 + \boldsymbol{U}_{i+n_1}, i = 1, \ldots, n_2$. Armed with this notation, we proceed to formulate the influence function expansion for sparse LDA.

We can see again, as in Section 4.5.2, that due to the special structure of the Sparse LDA estimator, it doesn't quite fall into the framework of Theorem 4.3.3. However, since the difference is only through adding two means — $\bar{\boldsymbol{X}}$ and $\bar{\boldsymbol{Y}}$, we can still can handle the asymptotics, as we demonstrate below.

Theorem 4.6.1. *Assume that $\lambda_{\min}(\mathbf{\Sigma}) > \delta > 0$, and that the two populations are coming from the same but shifted sub-Gaussian distribution as specified above. We further assume that the samples from the two populations are of comparable size $n_1 \asymp n_2 \asymp n$. Denote with s and $s_{\mathbf{v}}$ the sparsities of the vectors $\boldsymbol{\beta}^*$ and $\mathbf{v}^*$ correspondingly. Under the assumption that $\max(s_{\mathbf{v}}, s)\|\mathbf{v}^*\|_1(\|\boldsymbol{\beta}^*\|_1 \vee 1)\frac{\log d}{\sqrt{n}} = o(1)$, and large enough tuning parameters with $\lambda \asymp (\|\boldsymbol{\beta}^*\|_1 \vee 1)\sqrt{\frac{\log d}{n}}$ and $\lambda' \asymp \|\mathbf{v}^*\|_1\sqrt{\frac{\log d}{n}}$, we have the following asymptotic influence function expansion of the test statistic:*

$$n^{1/2}\widehat{S}(0, \widehat{\gamma}) = \frac{\mathbf{v}^{*T}}{n^{1/2}} \sum_{i=1}^{n} \left(\boldsymbol{U}_i^{\otimes 2}\boldsymbol{\beta}^* - (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) + \left[\frac{n}{n_1} I(i \leq n_1) - \frac{n}{n_2} I(i > n_1) \right] \boldsymbol{U}_i \right) + o_p(1).$$

Remark 4.6.2. *As before we have that $\mathbf{v}^* \geq 2K_{\boldsymbol{U}}^{-2}$ and hence we are guaranteed to have $\max(s_{\mathbf{v}}, s)(\|\boldsymbol{\beta}^*\|_1 \vee 1)\frac{\log d}{\sqrt{n}} = o(1)$.*

The proof of Theorem 4.6.1 can be found in Appendix C.4. Akin to the CLIME case we next discuss the following corollary:

Corollary 4.6.3. *Assume the same assumptions as in Theorem 4.6.1, and in addition assume that*

there exists $0 < \alpha < 1$ *such that* $n_1 - n\alpha = o(1)$*, and:*

$$\alpha \overbrace{\mathrm{Var}(\mathbf{v}^{*T}\boldsymbol{U}^{\otimes 2}\boldsymbol{\beta}^* + \alpha^{-1}\mathbf{v}^{*T}\boldsymbol{U})}^{V_1} + (1-\alpha)\overbrace{\mathrm{Var}(\mathbf{v}^{*T}\boldsymbol{U}^{\otimes 2}\boldsymbol{\beta}^* - (1-\alpha)^{-1}\mathbf{v}^{*T}\boldsymbol{U})}^{V_2}$$
$$\geq V_{\min}(\|\boldsymbol{\beta}^*\|_2^2\|\mathbf{v}^*\|_2^2 + \|\mathbf{v}^*\|_2^2) > 0. \tag{4.6.2}$$

Furthermore let $\frac{(s_{\mathbf{v}}s)^{3/2}}{n^{1/2}} = o(1)$*. Then we have:*

$$\frac{n^{1/2}}{\sqrt{\Delta}}\widehat{S}(0,\widehat{\boldsymbol{\gamma}}) \rightsquigarrow N(0,1), \quad \Delta = \alpha V_1 + (1-\alpha)V_2. \tag{4.6.3}$$

Remark 4.6.4. *Note that a sufficient condition for (4.6.2) to hold is condition (4.5.2) to hold. To see this first note that:*

$$\alpha V_1 + (1-\alpha)V_2 = \mathrm{Var}(\mathbf{v}^{*T}\boldsymbol{U}^{\otimes 2}\boldsymbol{\beta}^*) + \alpha^{-1}\mathbb{E}(\mathbf{v}^{*T}\boldsymbol{U})^2 + (1-\alpha)^{-1}\mathbb{E}(\mathbf{v}^{*T}\boldsymbol{U})^2.$$

Since we are assuming that $\mathbf{v}^{*T}\mathbb{E}\boldsymbol{U}^{\otimes 2}\mathbf{v}^* \geq \delta\|\mathbf{v}^*\|_2^2$*, we have:*

$$\alpha V_1 + (1-\alpha)V_2 \geq \mathrm{Var}(\mathbf{v}^{*T}\boldsymbol{U}^{\otimes 2}\boldsymbol{\beta}^*) + \delta(\alpha^{-1} + (1-\alpha)^{-1})\|\mathbf{v}^*\|_2^2.$$

Therefore if condition (4.5.2) holds with a constant $V'_{\min}$ *we have:*

$$\alpha V_1 + (1-\alpha)V_2 \geq \min(V'_{\min}, \delta(\alpha^{-1} + (1-\alpha)^{-1}))(\|\boldsymbol{\beta}^*\|_2^2\|\mathbf{v}^*\|_2^2 + \|\mathbf{v}^*\|_2^2).$$

As we saw in Remark 4.5.5, this implies that (4.6.2) holds for a multivariate normal distribution. Moreover, also showed that $\Delta \geq \delta(\alpha^{-1} + (1-\alpha)^{-1})\|\mathbf{v}^*\|_2^2 \geq \delta(\alpha^{-1} + (1-\alpha)^{-1})4K_{\boldsymbol{U}}^{-4} > 0$*.*

Below we propose an estimator of the variance Δ. Define:

$$\begin{aligned}\widehat{\Delta} = &\frac{1}{n}\sum_{i=1}^{n_1}\left(\widehat{\mathbf{v}}^T(\boldsymbol{X}_i - \bar{\boldsymbol{X}})^{\otimes 2}\widehat{\boldsymbol{\beta}}\right)^2 + \frac{1}{n}\sum_{i=1}^{n_1}\left(\frac{n}{n_1}\widehat{\mathbf{v}}^T(\boldsymbol{X}_i - \bar{\boldsymbol{X}})\right)^2 \\ &+ \frac{1}{n}\sum_{i=n_1+1}^{n}\left(\widehat{\mathbf{v}}^T(\boldsymbol{Y}_i - \bar{\boldsymbol{Y}})^{\otimes 2}\widehat{\boldsymbol{\beta}}\right)^2 + \frac{1}{n}\sum_{i=n_1+1}^{n}\left(\frac{n}{n_2}\widehat{\mathbf{v}}^T(\boldsymbol{Y}_i - \bar{\boldsymbol{Y}})\right)^2 \\ &- (\widehat{\mathbf{v}}^T(\bar{\boldsymbol{X}} - \bar{\boldsymbol{Y}}))^2.\end{aligned}$$

Proposition 4.6.5. *Under the same conditions as in Corollary 4.6.3, and the following additional assumptions:*

$$\begin{aligned}\|\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2\|_\infty^2\|\mathbf{v}^*\|_1 s_{\mathbf{v}}\lambda' &= o(1),\\ \lambda' s_{\mathbf{v}}\|\mathbf{v}^*\|_1\|\boldsymbol{\beta}^*\|_1\left(\lambda + \|\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2\|_\infty\right)\log(nd)\|\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2\|_\infty &= o(1),\\ \|\mathbf{v}^*\|_1\log(nd)s\lambda\|\boldsymbol{\beta}^*\|_1(1 + s_{\mathbf{v}}\lambda') &= o(1),\\ \|\boldsymbol{\beta}^*\|_1\left(\|\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2\|_\infty + \lambda\right)\lambda'(\sqrt{\log(nd)} + \|\boldsymbol{\mu}_1\|_\infty + \|\boldsymbol{\mu}_2\|_\infty) &= o(1),\\ \mathrm{Var}((\mathbf{v}^{*T}\boldsymbol{U})^2) = o(n), \quad \mathrm{Var}(\mathbf{v}^{*T}\boldsymbol{U}^{\otimes 2}\boldsymbol{\beta}^*) &= o(n),\end{aligned}$$

we have that $\widehat{\Delta} \to_p \Delta$.

Let $\widehat{U}_n = \frac{n^{1/2}}{\sqrt{\widehat{\Delta}}}\widehat{S}(0, \widehat{\boldsymbol{\gamma}})$. Combining the results from Proposition 4.6.5 and Theorem 4.6.1, we get the following for a fixed $t \in \mathbb{R}$:

Theorem 4.6.6. *Assume all assumptions in Proposition 4.6.5 and Theorem 4.6.1. Then the statistic* $\widehat{U}_n = \frac{n^{1/2}}{\sqrt{\widehat{\Delta}}}\widehat{S}(0, \widehat{\boldsymbol{\gamma}})$ *satisfies:*

$$\lim_{n\to\infty}|\mathbb{P}^*(\widehat{U}_n \le t) - \Phi(t))| = 0,$$

for any fixed $t \in \mathbb{R}$.

4.6.1 One-step Estimator and Confidence Intervals

It is easy to see that the one-step estimator, take the form:

$$\widetilde{\theta} = \widehat{\theta} - \frac{\widehat{\mathbf{v}}^T(\widehat{\mathbf{\Sigma}}_n\widehat{\beta} - (\bar{X} - \bar{Y}))}{\widehat{\mathbf{v}}^T(\widehat{\mathbf{\Sigma}}_n)_{*1}}. \tag{4.6.4}$$

We now formulate the following:

Corollary 4.6.7. *Assume the same conditions as in Corollary 4.6.3. Then we have*

$$\frac{n^{1/2}}{\sqrt{\Delta}}(\widetilde{\theta} - \theta^*) \rightsquigarrow N(0, 1),$$

where Δ is defined as in (4.6.3).

Proof. The proof is identical to the proof of Corollary 4.5.14, so we omit it. □

Remark 4.6.8. *Under the assumptions of Proposition 4.6.5, we can use $\widehat{\Delta}$ as a consistent estimate to construct confidence intervals in practice.*

4.7 Stationary Vector Autoregressions

In a recent paper by Han et al. [29], the authors proposed a new estimator for transition matrices in high dimensional vector autoregressions. The idea of their estimator is similar to the CLIME idea, and hence it fits in the framework discussed throughout this chapter. In this section we complement their theory with developing inferential procedures.

For convenience we will only consider the case of a lag 1 models. As mentioned in Han et al. [29], lag p models can be accommodated in the framework of lag 1 models, and thus we are not losing any generality in doing so. We review some basic notations for autoregressive processes below.

Let $(\boldsymbol{X}_t)_{t=-\infty}^{\infty}$ be a stationary sequence of 0 mean random vectors in $\mathbb{R}^d$ with covariance matrix $\boldsymbol{\Sigma}$. The sequence $(\boldsymbol{X}_t)_{t=-\infty}^{\infty}$ is said to follow a lag 1 autoregressive model iff:

$$\boldsymbol{X}_t = A^T \boldsymbol{X}_{t-1} + \boldsymbol{Z}_t, \quad t \in \mathbb{Z}.$$

The matrix A is called transition matrix. It is further assumed that the noise vectors $\boldsymbol{Z}_t$ are independent and identically distributed with $\boldsymbol{Z}_t \sim N(0, \boldsymbol{\Psi})$. Moreover it is assumed that $\boldsymbol{Z}_t$ is independent of the history $(\boldsymbol{X}_s)_{s<t}$. Under the additional assumption that $\det(I_d - A^T z) \neq 0$ for all $z \in \mathcal{C}$ with $|z| \leq 1$, it can be shown that the $\boldsymbol{\Psi}$ can be selected so that the process is stationary, i.e. for all t: $\boldsymbol{X}_t \sim N(0, \boldsymbol{\Sigma})$.

Let $\Sigma_i\{(\boldsymbol{X}_t)\} = \mathrm{Cov}(\boldsymbol{X}_0, \boldsymbol{X}_i)$, so that $\Sigma_0\{(\boldsymbol{X}_t)\} = \boldsymbol{\Sigma}$. A simple calculation in the lag 1 case leads to the Yule-Walker Equation below:

$$\Sigma_i\{(\boldsymbol{X}_t)\} = \Sigma_0\{(\boldsymbol{X}_t)\} A^i.$$

A trivial consequence of the latter equation is that:

$$A = (\Sigma_0\{(\boldsymbol{X}_t)\})^{-1} \Sigma_1\{(\boldsymbol{X}_t)\}.$$

Realizing this, Han et al. [29] propose to solve the following optimization problem in the high-dimensional setting:

$$\widehat{A} = \underset{M \in \mathbb{R}^{d \times d}}{\operatorname{argmin}} \sum_{jk} |M_{jk}|, \text{ subject to } \|S_0 M - S_1\|_{\max} \leq \lambda, \tag{4.7.1}$$

where $\lambda > 0$ is a tunning parameter, $S_0 = \frac{1}{T} \sum_{t=1}^{T} \boldsymbol{X}_t^{\otimes 2}$ and $S_1 = \frac{1}{T-1} \sum_{t=1}^{T-1} \boldsymbol{X}_t \boldsymbol{X}_{t+1}^T$ are estimates of Σ_0 and Σ_1 correspondingly and T is the number of observations of the time series.

The above formulation bares similarity to the CLIME procedure, and similarly to CLIME it can be decomposed into subproblems for each column of A. Let $\boldsymbol{\beta}^* = A_{*m}$ be the m^{th} column of A. If one is interested in only the estimate of $\boldsymbol{\beta}^*$ from (4.7.1), instead of solving the whole problem one can only solve the corresponding subproblem:

$$\widehat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta}\in\mathbb{R}^d}{\operatorname{argmin}} \|\boldsymbol{\beta}\|_1, \text{ subject to } \|S_0\boldsymbol{\beta} - S_{1,*m}\|_{\max} \leq \lambda, \tag{4.7.2}$$

In [29], the authors showed that procedure (4.7.1) consistently estimates A under certain sparsity assumptions on A. Along the way they developed concentration bounds for S_0 and S_1, which we use in the present development. In this section we propose a testing procedure for testing $H_0 : A_{1m} = 0$ vs $H_A : A_{1m} \neq 0$, where as usual 1 is selected without loss of generality. To this end we consider the following optimization problem:

$$\widehat{\mathbf{v}} = \min_{\mathbf{v}\in\mathbb{R}^d} \|\mathbf{v}\|_1, \text{ subject to } \|\mathbf{v}^T S_0 - \mathbf{e}\|_{\max} \leq \lambda',$$

where $\widehat{\mathbf{v}}$ is intended as an estimate of $\mathbf{v}^{*T} = \Sigma_{0,*1}$. Next we define the test statistic:

$$\widehat{S}(\boldsymbol{\beta}) = \widehat{\mathbf{v}}^T(S_0\boldsymbol{\beta} - S_1).$$

Note that this framework differs from the general procedures developed in Section 4.3, in that there is dependency between the observations, and furthermore $S_0 v - S_{1,*m} = 0$ is not a typical estimating equation described in (4.2.1). Nevertheless using similar ideas we can make the theory go through in this case.

To this end we define several quantities which play important role in our analysis. Let $M_d \in \mathbb{R}$

be a constant which is allowed to scale with (T, d). We next define a class of matrices:

$$\mathcal{M}(s, M_d) := \left\{ M \in \mathbb{R}^{d\times d} : \max_{1\leq j\leq d} \|M_{*j}\|_0 \leq s, \|M\|_1 \leq M_d \right\}$$

From now on we will assume that the transition matrix $A \in \mathcal{M}(s, M_d)$. Next we define two important complexity measures:

$$K_d(\Sigma_0, A) := \frac{32\|\Sigma_0\|_2 \max_j(\Sigma_{0,jj})}{\min_j(\Sigma_{0,jj})(1 - \|A\|_2)}, \quad \widetilde{K}_d(\Sigma_0, A) := K_d(\Sigma_0, A)(2M_d + 3).$$

We will furthermore assume that the vector $\mathbf{v}^*$ is sparse with $\|\mathbf{v}^*\|_0 = s_{\mathbf{v}}$.

We then proceed to formulate an influence function expansion.

Theorem 4.7.1. *Set* $\lambda = \widetilde{K}_d(\Sigma_0, A)\sqrt{\frac{\log d}{T}}$ *and* $\lambda' = \frac{K_d(\Sigma_0,A)}{2}\|\Sigma_0^{-1}\|_1 \left(\sqrt{\frac{6\log d}{T}} + 2\sqrt{\frac{1}{T}}\right)$. *Assume that*

$$\lambda = o(1), \quad \lambda' = o(1), \quad \sqrt{T}\max(s_{\mathbf{v}}, s)\|\Sigma_0^{-1}\|_1\lambda'\lambda = o(1). \tag{4.7.3}$$

If additionally $T \geq 6\log(d+1)$ *and* $d \geq 8$ *we have the following:*

$$\sqrt{T}\widehat{S}(0, \widehat{\boldsymbol{\gamma}}) = \sqrt{T}\mathbf{v}^{*T}(S_0\boldsymbol{\beta}^* - S_{1,*m}) + o_p(1).$$

The proof of Theorem 4.7.1 can be found in Appendix C.5. Next we provide a weak convergence result.

Theorem 4.7.2. *Assume the assumptions from Theorem 4.7.1. Furthermore assume the following*

regularity conditions:

$$\mathbf{\Psi}_{mm}\mathbf{v}^{*T}\Sigma_0\mathbf{v}^{*T} \geq C' > 0, \quad \frac{\boldsymbol{\beta}^{*T}\Sigma_0\boldsymbol{\beta}^*}{\mathbf{\Psi}_{mm}} = o(T), \quad \frac{\lambda' M_d}{\sqrt{\mathbf{\Psi}_{mm}\mathbf{v}^{*T}\Sigma_0\mathbf{v}^{*T}}} = o(\sqrt{T}), \tag{4.7.4}$$

$$\frac{\|\mathbf{v}^*\|_1^2}{\mathbf{v}^{*T}\Sigma_0\mathbf{v}^*}\frac{\lambda'}{\|\Sigma_0^{-1}\|_1} = o(1). \tag{4.7.5}$$

Then we have:

$$\frac{T^{1/2}\widehat{S}(0,\widehat{\boldsymbol{\gamma}})}{\sqrt{\mathbf{\Psi}_{mm}\mathbf{v}^{*T}\Sigma_0\mathbf{v}^*}} \rightsquigarrow N(0,1).$$

As before, in practice we need a consistent estimate of the variance $\Delta = \mathbf{\Psi}_{mm}\mathbf{v}^{*T}\Sigma_0\mathbf{v}^*$. For that purpose we consider the following:

Proposition 4.7.3. *Let* $\widehat{\Delta} = (S_{0,mm} - \widehat{\boldsymbol{\beta}}^T S_0 \widehat{\boldsymbol{\beta}})(\widehat{\mathbf{v}}^T S_0 \widehat{\mathbf{v}})$. *Assume the notation and assumptions of Theorem 4.7.1. Under the following additional assumptions:*

$$\lambda' \max(\|\Sigma_0^{-1}\|_1^{-1}, \|\Sigma_0^{-1}\|_1) = o(1), \quad 4s\|\Sigma_0^{-1}\|_1\lambda M_d \max(\|\Sigma_0\|_{\max}, 1) = o(1) \tag{4.7.6}$$

$$\lambda \max(M_d, 1) = o(1), \quad M_d K_d(\Sigma_0, A)\left(\sqrt{\frac{3\log d}{T}} + \sqrt{\frac{2}{T}}\right) = o(1), \tag{4.7.7}$$

$$(\lambda')^2 s_{\mathbf{v}}\|\Sigma_0^{-1}\|_1 = o(1), \tag{4.7.8}$$

we have $\widehat{\Delta} \to_p \Delta$.

The proof of Proposition 4.7.3 is deferred to Appendix C.5. It enables testing in practical setting. We have the following Corollary which we formulate without a proof:

Corollary 4.7.4. *Assume all assumptions in Theorems 4.7.1 and 4.7.2 and Proposition 4.7.3. Then the statistic* $\widehat{U}_T = \frac{T^{1/2}}{\sqrt{\widehat{\Delta}}}\widehat{S}(0,\widehat{\boldsymbol{\gamma}})$ *satisfies:*

$$\lim_{n\to\infty} |\mathbb{P}^*(\widehat{U}_n \leq t) - \Phi(t))| = 0,$$

for any fixed $t \in \mathbb{R}$.

4.7.1 One-step Estimator and Confidence Intervals

It can be seen that the one-step estimator take the following form:

$$\widetilde{\theta} = \widehat{\theta} - \frac{\widehat{\mathbf{v}}^T(S_0\widehat{\boldsymbol{\beta}} - S_{1,*m})}{\widehat{\mathbf{v}}^T S_{0,*1}}$$

Next we can formulate the following:

Corollary 4.7.5. *Assume the same conditions as in Corollary 4.7.4. Then we have*

$$\frac{n^{1/2}}{\sqrt{\Delta}}(\widetilde{\theta} - \theta^*) \rightsquigarrow N(0,1).$$

Remark 4.7.6. *Under the assumptions of Proposition 4.7.3, we can substitute* Δ *with* $\widehat{\Delta}$ *not changing the weak convergence to enable confidence interval construction in practical settings.*

4.8 Quasi-Likelihood

In this section we consider an extension to quasi-likelihood equations. Let us observe n iid samples $(Y_i, \boldsymbol{X}_i)$. A general quasi-likelihood equation is based on two moments $\mathbb{E}y = \boldsymbol{\mu}$, and $\mathrm{Var}(y) = v(\boldsymbol{\mu})a^{-1}(\phi)$, and on a link function g which links the mean to the linear component $g(\boldsymbol{\mu}) = \boldsymbol{X}^T\boldsymbol{\beta}$. For simplicity we will consider the special case when $g'(\boldsymbol{\mu}) = v(\boldsymbol{\mu})^{-1}$ and $a(\phi) = 1$ corresponding to the case where the link function g is the canonical link. Notice that canonical links are by definition strictly increasing and let $f = g^{-1}$ denote the inverse link function such that $\boldsymbol{\mu} = f(\boldsymbol{X}^T\boldsymbol{\beta})$. We will also assume that f is continuously differentiable. In this special case the

quasi-likelihood equation becomes:

$$n^{-1}\sum_{i=1}^{n}(f(\boldsymbol{X}_i^T\boldsymbol{\beta})-Y_i)\boldsymbol{X}_i=0.$$

Hence according to our framework we will determine the estimate of $\boldsymbol{\beta}$ through:

$$\widehat{\boldsymbol{\beta}}=\text{argmin}\,\|\boldsymbol{\beta}\|_1,\ \text{subject to}\left\|n^{-1}\sum_{i=1}^{n}(f(\boldsymbol{X}_i^T\boldsymbol{\beta})-Y_i)\boldsymbol{X}_i\right\|_\infty\leq\lambda.$$

As usual we are interested in testing $H_0:\theta=0$ vs $H_A:\theta\neq 0$, where θ is the first component of $\boldsymbol{\beta}$. We will use the statistic $\widehat{S}(\boldsymbol{\beta})=n^{-1}\widehat{\mathbf{v}}^T\sum_{i=1}^{n}(f(\boldsymbol{X}_i^T\boldsymbol{\beta})-Y_i)\boldsymbol{X}_i$. Here $\widehat{\mathbf{v}}$, based on the following program:

$$\widehat{\mathbf{v}}=\text{argmin}\,\|\mathbf{v}\|_1,\ \text{subject to}\left\|n^{-1}\sum_{i=1}^{n}\mathbf{v}^T f'(\boldsymbol{X}_i^T\widehat{\boldsymbol{\beta}})\boldsymbol{X}_i^{\otimes 2}-\mathbf{e}\right\|_\infty\leq\lambda'.$$

For brevity define $\boldsymbol{\Sigma}_{n,W}=n^{-1}\sum_{i=1}^{n}f'(\boldsymbol{X}_i^T\boldsymbol{\beta}^*)\boldsymbol{X}_i^{\otimes 2}$, and the population version $\boldsymbol{\Sigma}_W=\mathbb{E}f'(\boldsymbol{X}^T\boldsymbol{\beta}^*)\boldsymbol{X}^{\otimes 2}$

To this end we formulate:

Assumption 4.8.1. *Assume that the covariates are bounded, i.e. there exist constants $K,K'>0$ such that $\|\boldsymbol{X}\|_\infty\leq K$, $|\boldsymbol{\beta}^{*T}\boldsymbol{X}|\leq K$ and $|Y-\boldsymbol{\beta}^{*T}\boldsymbol{X}|\leq K'$ hold with probability 1. Furthermore assume that for any $x,y\in[-2K,2K]$ we have $|f'(x)-f'(y)|\leq C|x-y|$ and $|f'(x)|<C$ for some fixed $C>0$.*

It is clear that these assumptions are stronger than the sub-Gaussian assumptions we have made so far. They can be relaxed to sub-Gaussian assumptions, but we believe that adopting Assumption 4.8.1 makes the presentation cleaner, while preserving the innate difficulty of the problem. Consider the following:

Theorem 4.8.2. *Let Assumption 4.8.1 hold and in addition assume* $\mathbf{\Sigma}_W \geq \delta > 0$. *Set* $\lambda = 2K'K\sqrt{\frac{\log d}{n}}$, *and*

$$\lambda' = \|\mathbf{v}^*\|_1 \left(\frac{8(2+CK^3)\lambda s}{\lambda_{\min}(\mathbf{\Sigma}_W)} + \sqrt{6}CK^2\sqrt{\frac{\log d}{n}} \right).$$

Further, assume that the following relationships hold:

$$\sqrt{n}\lambda'\lambda \max(s_{\mathbf{v}}, s) = o(1),\ \lambda s^2 = o(1),\ \lambda = o(1),\ \lambda' = o(1).$$

Then we have the influence function expansion:

$$\sqrt{n}\widehat{S}(\widehat{\boldsymbol{\beta}}_0) = \sqrt{n}S(\boldsymbol{\beta}^*) + o_p(1).$$

Remark 4.8.3. *Note that the assumptions in this theorem are more stringent than the theorem for the linear model. The strengthening is required because of the dependence on* $\boldsymbol{\beta}$ *of the Hessian matrix* $-n^{-1}\sum_{i=1}^n \boldsymbol{X}_i^{\otimes 2} f'(\boldsymbol{X}_i^T\widehat{\boldsymbol{\beta}})$.

The proof of Theorem 4.8.2 can be found in Appendix C.6. We next provide an appropriate standardization to show the CLT. Denote with $\Delta := \mathbf{v}^{*T}\mathbf{\Sigma}_W\mathbf{v}^*$. We have:

Corollary 4.8.4. *Under the same assumptions as in Theorem 4.8.2, and in addition* $s_{\mathbf{v}}^{3/2}/n^{1/2} = o(1)$, *we have that:*

$$\frac{n^{1/2}}{\sqrt{\Delta}}\widehat{S}(0, \widehat{\boldsymbol{\gamma}}) \rightsquigarrow N(0,1).$$

Next we provide consistent estimates for Δ. In this case, one obvious candidate for such statistic is $\widehat{\Delta}_1 = \widehat{\mathbf{v}}_1$. Furthermore we can use $\widehat{\Delta}_2 = n^{-1}\sum_{i=1}^n (\widehat{\mathbf{v}}^T\boldsymbol{X}_i)^2 f'(\boldsymbol{X}_i^T\widehat{\boldsymbol{\beta}})$ or $\widehat{\Delta}_3 = n^{-1}\sum_{i=1}^n (\widehat{\mathbf{v}}^T\boldsymbol{X}_i)^2 (Y_i - f(\boldsymbol{X}_i^T\widehat{\boldsymbol{\beta}}))^2$.

Proposition 4.8.5. *Assume:*

1. *the assumptions of Lemma C.6.6 and* $\lambda' s_{\mathbf{v}} = o(1)$,

2. *1. and* $\lambda' \|\mathbf{v}^*\|_1 = o(1)$,

3. *1.,* $\lambda s \|\mathbf{v}^*\|_1^2 = o(1)$ *and* $\|\mathbf{v}^*\|_1^2 \sqrt{\frac{\log d}{n}} = o(1)$.

Under assumption i. for $i = 1, 2, 3$*, we have:*

$$\widehat{\Delta}_i \to_p \Delta.$$

The proof of Proposition 4.8.5 can be found in Appendix C.6. Given Proposition 4.8.5 the following Theorem is an implication of Slutsky's theorem:

Theorem 4.8.6. *Assume all assumptions in Corollary 4.8.4, and construct estimates of* $\Delta - \widehat{\Delta}_i$ $i = 1, 2, 3$ *based on Proposition 4.8.5 under their corresponding conditions. Then the statistics* $\widehat{U}_n^i = \frac{n^{1/2}}{\widehat{\Delta}_i} \widehat{S}(0, \widehat{\gamma})$ *satisfy:*

$$\lim_{n \to \infty} |\mathbb{P}^*(\widehat{U}_n^i \leq t) - \Phi(t))| = 0, \;\; i = 1, 2, 3.$$

Remark 4.8.7. *Note that the assumptions of Lemma C.6.6 are already implied by the assumptions of Corollary 4.8.4, and hence a variance estimator which requires no new assumptions is* $\widehat{\Delta}_1$.

4.8.1 One-Step Estimator and Confidence Intervals

It can be seen that the one-step estimator take the following form:

$$\widetilde{\theta} = \widehat{\theta} - \frac{\widehat{\mathbf{v}}^T \sum_{i=1}^n \boldsymbol{X}_i (f(\boldsymbol{X}_i^T \widehat{\boldsymbol{\beta}}) - Y_i)}{\widehat{\mathbf{v}}^T [\sum_{i=1}^n \boldsymbol{X}_i^{\otimes 2} f'(\boldsymbol{X}_i^T \widehat{\boldsymbol{\beta}})]_{*1}}.$$

Next we can formulate the following:

Corollary 4.8.8. *Assume the same conditions as in Corollary 4.8.4. Furthermore let:*

$$n^{1/2}\lambda'\lambda s = o(1).$$

Then we have:

$$\frac{n^{1/2}}{\sqrt{\Delta}}(\widetilde{\theta} - \theta^*) \rightsquigarrow N(0,1).$$

The proof Corollary 4.8.8 can be found in Appendix C.6.

Remark 4.8.9. *Under the assumptions of Proposition 4.8.5, we can substitute Δ with $\widehat{\Delta}_i, i = 1,2,3$ not changing the weak convergence to enable confidence interval construction in practical settings.*

4.9 Numerical Studies

In this section we present our numerical evidence in support to our theoretical claims.

4.9.1 Linear Model

In this section we compare our results to two existing methods – the desparsity[80] and the debias[36] methods. We stress the fact that these two methods, are different from the one we are currently suggesting, in that both of them are based on using the LASSO as initial estimator rather than solving the estimating equation with a Dantzig Selector.

Our simulation setup is the following: we generate $n = 150$ observations $X \sim N(0, \mathbf{\Sigma_X})$, where $\mathbf{\Sigma_X}$ is a Toeplitz matrix with $\mathbf{\Sigma_X}_{,ij} = \rho^{|i-j|}, i,j = 1,\ldots,d$. We consider several scenarios for the correlation $\rho \in \{0.25, 0.4, 0.6\}$. Furthermore, we have 3 possible values of the dimension $d = 100, 200, 500$. To asses the size of the three procedures we generated $\boldsymbol{\beta}^*$ under two settings. In the first setting $\boldsymbol{\beta}^*$was held fixed $\boldsymbol{\beta}^* = (1,1,1,\underbrace{0,\ldots,0}_{d-3})^T$, and for the second setting $\beta^* =$

$(U_1, U_2, U_3, \underbrace{0, \ldots, 0}_{d-3})^T$ where $U_i \sim U([0,2]), i = 1,2,3$. The former setting is labeled as "Dirac" and the latter as "Uniform" in Table 4.1 below. The outcome value $y = X^T\boldsymbol{\beta}^* + \varepsilon$, where $\varepsilon \sim N(0,1)$. Each of the simulations is repeated 500 times.

The tuning parameter λ was selected by a 10-fold cross validation. The parameter λ' was manually set to $\frac{1}{2}\sqrt{\frac{\log d}{n}}$. We discovered that the test is robust with respect to the choice of λ'.

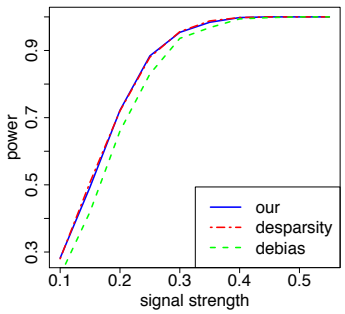
A summary of the size results can be found in the table below, for the test $H_0 : \boldsymbol{\beta}_1^* = 1$ vs $H_A : \boldsymbol{\beta}_1^* \neq 1$ in the first setting and $H_0 : \boldsymbol{\beta}_1 = \boldsymbol{\beta}_1^*$ vs $H_A : \boldsymbol{\beta}_1 \neq \boldsymbol{\beta}_1^*$ in the second one.

Table 4.1: Size in the Linear Model

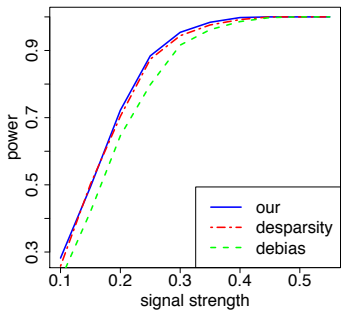
		Dirac			Uniform		
d	method	$\rho = 0.25$	$\rho = 0.4$	$\rho = 0.6$	$\rho = 0.25$	$\rho = 0.4$	$\rho = 0.6$
	our	0.054	0.056	0.056	0.056	0.052	0.062
100	desparisty	0.056	0.052	0.052	0.052	0.054	0.046
	debias	0.06	0.052	0.06	0.052	0.054	0.06
	our	0.048	0.058	0.048	0.05	0.052	0.054
200	desparisty	0.038	0.064	0.048	0.044	0.064	0.054
	debias	0.05	0.054	0.058	0.05	0.048	0.060
	our	0.058	0.052	0.046	0.042	0.058	0.05
500	desparisty	0.052	0.058	0.054	0.046	0.058	0.056
	debias	0.058	0.048	0.056	0.054	0.06	0.036

For power analysis we used $\boldsymbol{\beta}^* = (\xi, \xi, \xi, \underbrace{0, \ldots, 0}_{d-3})^T$ where ξ took values in the interval $[.1, .55]$.

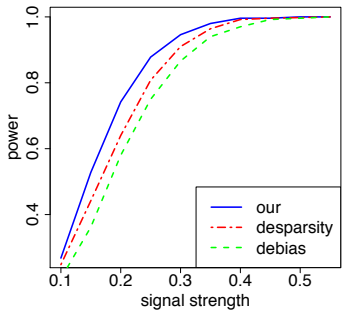
Figure 4.1: Power Comparisons for the Linear Models



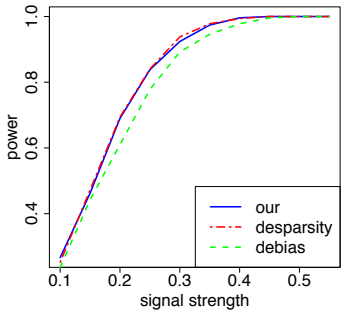
$\rho = 0.25, d = 100$



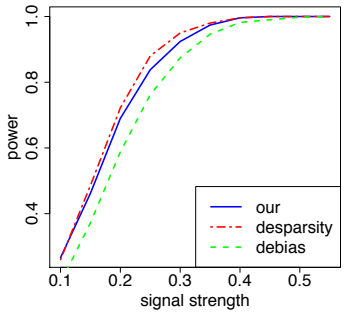
$\rho = 0.4, d = 100$



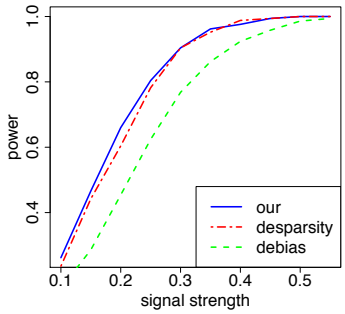
$\rho = 0.6, d = 100$



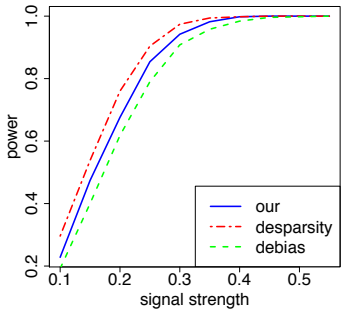
$\rho = 0.25, d = 200$



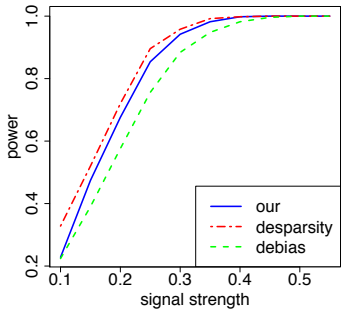
$\rho = 0.4, d = 200$



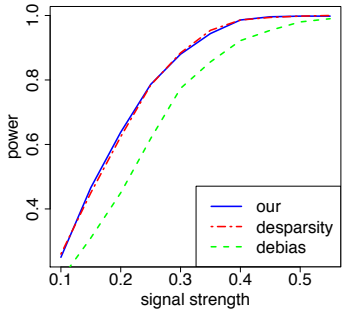
$\rho = 0.6, d = 200$



$\rho = 0.25, d = 500$



$\rho = 0.4, d = 500$



$\rho = 0.6, d = 500$

We tested $H_0 : \beta_1^* = 0$ vs $H_A : \beta_1^* \neq 0$ and assessed the power for the three algorithms. The

power plots can be found in figure 4.1.

As we can see from the power plots, our proposed test statistic performs very similarly to the ones proposed in the two other papers in the linear model. This is to be expected as all of the three statistics are asymptotically optimal and hence the powers should be equivalent.

4.9.2 Graphical Models

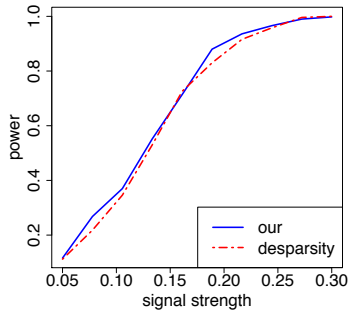
In this section we compare our CLIME-based procedure to the custom desparsifying algorithm defined by Jankova and van de Geer [35], based on the graphical LASSO. We took one of the examples considered in their paper, and transformed it into a power comparison. We considered a tridiagonal precision matrix $\mathbf{\Omega}$ with $\mathbf{\Omega}_{ii} = 1, i = 1, \ldots, d$ and $\mathbf{\Omega}_{i,i+1} = \mathbf{\Omega}_{i+1,i} = 0.3$ for $i = 1, \ldots, d-1$.

We considered $d = 80$, and $n = 250$ as in Jankova and van de Geer [35]. The λ tunning parameter was set equal to $0.5\sqrt{\frac{\log d}{n}}$ in both algorithms, while we discovered that the $\lambda' = \sqrt{\frac{\log d}{n}}$ for our method gave a more precise size results, although the choice was fairly robust. Below we present the size results of testing $\mathbf{\Omega}_{12} = 0$ under this scenario.

	CLIME EE	desparsity
$d = 80$	0.050	0.056

Below we are attaching the power plots under the same scenario, where we are ranging the signal strength $\rho \in [0.05, 0.3]$.

Figure 4.2: CLIME EE vs Graphical Lasso desparsity



In another experiment we generated data through the following procedure, inspired by Liu et al.[53]. The latent generalized concentration matrix $\mathbf{\Omega}^*$ was generated in the same way as in our previous example, but then was normalized so that $\mathbf{\Sigma}^* = \mathbf{\Omega}^{*-1}$, satisfies $\text{diag}(\mathbf{\Sigma}^*) = 1$. Then a normally distributed data was generated through $\boldsymbol{X}_i \sim N(0, \mathbf{\Sigma}^*), i = 1, \ldots, n$, and was transformed through the following marginal transformations.

Definition 4.9.1 (Symmetric Power Transformation, Liu et al.[53])**.** *Let f be:*

$$f(t) = \text{sign}(t)|t|^\alpha,$$

where $\alpha > 0$ is a parameter of the transformation. We then define the power transformation of the j^{th} dimension as:

$$\boldsymbol{Z}^j = g_j(\boldsymbol{X}^j) = \frac{f(\boldsymbol{X}^j)}{\sqrt{\int f^2(t)\phi\,(t)\,dt}}.$$

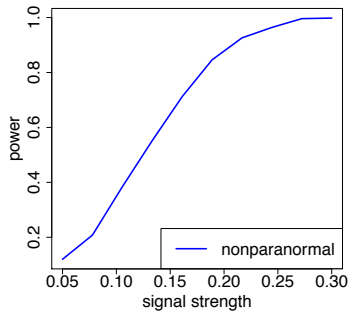
These transformations are designed to preserve the marginal mean and standard deviation. In our experiment we used a value of $\alpha = 5$. The observed data consistent of $\boldsymbol{Z}_i, i = 1, \ldots, n$. We assessed the size for the three procedures — the desparsity procedure based on Graphical LASSO, the CLIME based procedure and the nonparanormal procedure with CLIME. We set $\rho = 0.3$ and

tested $H_0 : \mathbf{\Omega}_{12} = \mathbf{\Omega}^*_{12}$ vs $H_A : \mathbf{\Omega}_{12} \neq \mathbf{\Omega}^*_{12}$. As expected the former two procedures could not give a correct size despite our efforts to select different ranges for the tuning parameters, as they are designed to test the parameter for the covariance of $\boldsymbol{Z}$ which has a different covariance structure than $\boldsymbol{X}$, and is not even coming from a sub-Gaussian distribution. The nonparanormal procedure with CLIME however performed quite well. Below we summarize the size results:

	CLIME EE	desparsity	nonparanormal CLIME EE
$d = 80$	–	–	0.050

The tuning parameters were selected in exactly the same way as we selected the tunning parameters for the CLIME testing procedure in our previous example. Below we present a power plot of the nonparanormal procedure.

Figure 4.3: Nonparanormal CLIME EE Power



We varied the signal $\rho \in [0.05, 0.3]$ range and tested for $H_0 : \mathbf{\Omega}_{12} = 0$ vs $H_A : \mathbf{\Omega}_{12} \neq 0$. As we can see from the power plot the nonparanormal with CLIME procedure performs quite well in this setting.

4.10 Discussion

In this chapter we proposed a generic procedure for testing linear Z estimators in a high-dimensional setting. We provided a general framework, and showed several important applications including in Linear Models and Graphical Models. We demonstrated through simulations, that our framework performs as well as previously suggested algorithms, but has the advantage of having a broader scope and covering many applications.

Much remains to be done in the current framework. In our future work, we will consider handling models with missing data and/or sampling bias, and extending our testing procedure to the non-linear case. Moreover, we plan on extending the one-dimensional testing to the multi-dimensional case. We note that the latter extension is not trivial. One possible approach is to use the multiplier bootstrap.

One never notices what has been done; one can only see what remains to be done.

Marie Curie

5 Conclusion

In this dissertation we discussed three important problems — classification, variable selection and statistical inference.

In Chapter 2 we expanded existing classes of loss functions that achieve Fisher consistency, and showed that non-convex loss functions should not be excluded from consideration for multi-class classification. Via simulations we showed that such non-convex losses can have similar and in cases better performances compared to convex loss functions which are typically used in practice, such as

the exponential loss and the logistic losses. We proposed a generic boosting algorithm, which can be used with any loss function from our class of relaxed Fisher consistent losses. We proved that this boosting algorithm converges to the global minimum, when the loss function is convex at a geometric rate. In terms of future directions, a lot remains to be done. We conjecture that in cases with non-convex loss functions, the boosting algorithm will have geometric rate of convergence to local minimizers, and that furthermore such local minimizers can be used to consistently recover the Bayes rule provided that the classifier bag is rich enough. Our Fisher consistency results strongly suggest that such a statement indeed holds, and we anticipate that it can be established with the help of tools from empirical process theory.

Chapter 3 discussed the behavior of sliced inverse regression in a high-dimensional setting. We demonstrated that diagonal thresholding and semidefinite programming can be used with SIR to recover the support of single index models with uncorrelated Gaussian predictors. We also derived a lower bound on the sample size in terms of the sparsity of the signal and the ambient dimension, of any algorithm which recovers the support with high-probability. Our results indicated that this lower bound is achieved (up to a constant) by the DT and SDP algorithms, and we backed up these theoretical claims with thorough simulations. To the best of our knowledge, this phase transition phenomenon has not been previously observed in the literature. Moreover, in a slightly more restrictive setting, when the predictor is correlated with the outcome, we showed that covariance thresholding can also achieve such an optimal sample size when the predictors are coming from a standard Gaussian ensemble. We also addressed the question of what can we do when the predictor matrix has a correlated Gaussian distribution, by showing that a linear regression LASSO can enjoy optimal sample size provided that the signal and the covariance satisfy certain sufficient conditions. This discussion leaves us with many standing questions that remain to be explored. One such question is: can we approach the general covariance problem directly in terms of the SIR algorithm? A solution could be estimating the covariance at a "good enough" rate and using the two algorithms we

discussed. Furthermore, can we construct direct non-convex penalization approaches to recover the support without the restrictive irrepresentable condition on the covariance matrix that the LASSO requires? Another important question is how to generalize our work for multi-index model.

In Chapter 4 we presented a novel framework for inference in high-dimensional estimating equations. We used our framework to equip many popular high-dimensional estimating procedures such as the Dantzig Selector, CLIME and LDP with inferential frameworks. Throughout this chapter, our theory focused on testing a one-dimensional component of the parameter of the estimating equation. Our framework can trivially be extended to cases where we are interested in testing finitely many components of the parameter of interest. A more interesting generalization, would be an extension allowing for the number of parameters to scale (at even exponential rates) with the sample size. Such an extension might indeed be possible, with the help of recent results on conditional multiplier central limit theorems. Furthermore, we have failed to address how would the properties of our inferential procedure change in cases of model mis-specification. In another train of thought, it will be interesting to consider different assumptions instead of the sparsity of the covariance operator that we are currently assuming.

We hope to address all of the open questions listed above in our future work. With this the dissertation is concluded.

A

Proofs for Chapter 2

Lemma A.0.1. *Assumption (2.2.5) implies that the function $\phi(g^{-1}(z))$ is continuously differentiable and convex for all $z \in g(S)$.*

Proof of Lemma A.0.1. Set $z := g(x), z' := g(x')$ in (2.2.5). When $x, x' \in S$ we have $z, z' \in g(S)$ and vice versa. Now (2.2.5) can be rewritten as:

$$\phi(g^{-1}(z)) - \phi(g^{-1}(z')) \geq (z - z')k(g^{-1}(z')). \tag{A.0.1}$$

Changing the roles of z and z' and using the fact that both $z, z' \in g(S)$ we obtain:

$$\phi(g^{-1}(z')) - \phi(g^{-1}(z)) \geq (z' - z)k(g^{-1}(z)).$$

The above two inequalities give that for any $z \neq z', z, z' \in g(S)$ we have:

$$\min\{k(g^{-1}(z)), k(g^{-1}(z'))\} \leq \frac{\phi(g^{-1}(z')) - \phi(g^{-1}(z))}{z' - z} \leq \max\{k(g^{-1}(z)), k(g^{-1}(z'))\}. \tag{A.0.2}$$

By the continuity of k and g we have that the composition $k(g^{-1}(\cdot))$ is also continuous. Taking the limit $z' \to z$ in (A.0.2) shows that the function $\phi(g^{-1}(z))$ is differentiable on $g(S)$ with a continuous derivative equal to $k(g^{-1}(z))$. Now the convexity of $\phi(g^{-1}(z))$ follows from (A.0.1).

□

Proof of Theorem 2.2.1. To show that $H_\phi(\widehat{F}_j)w_j = \mathcal{C}$ for some $\mathcal{C} < 0$, define $\Omega = \{\boldsymbol{F} = (F_1, \ldots, F_n) : F_j \in S, j = 1, ..., n\}$, where recall that $S = \{z \in \mathbb{R} : k(z) \leq 0\}$. From (2.2.5),

$$\sum_{j=1}^{n} \phi(\widehat{F}_j)w_j \geq \sum_{j=1}^{n} \phi(F_j)w_j + \sum_{j=1}^{n} \{g(\widehat{F}_j) - g(F_j)\}k(F_j)w_j \quad \text{for any } \boldsymbol{F} \in \Omega. \tag{A.0.3}$$

Since $\widehat{\boldsymbol{F}}$ minimizes $\sum_{j=1}^{n} \phi(F_j)w_j$, (A.0.3) implies that

$$\sum_{j=1}^{n} g(F_j)k(F_j)w_j \geq \sum_{j=1}^{n} g(\widehat{F}_j)k(F_j)w_j \quad \text{for any } \boldsymbol{F} \in \Omega. \tag{A.0.4}$$

For any given constant $\mathcal{C} < 0$, let $\widetilde{F}_{\mathcal{C}j}$ be the solution to $g(\widehat{F}_j)k(\widetilde{F}_{\mathcal{C}j})w_j = \mathcal{C}$ or equivalently $\widetilde{F}_{\mathcal{C}j} = k^{-1}[\mathcal{C}/\{g(\widehat{F}_j)w_j\}]$. Obviously $\widetilde{\boldsymbol{F}}_{\mathcal{C}} \in \Omega$ for all $\mathcal{C} < 0$. We next show that there exists $\mathcal{C}_0 < 0$ such that $\prod_{j=1}^{n} g(\widetilde{F}_{\mathcal{C}_0 j}) = \prod_{j=1}^{n} g(k^{-1}[\mathcal{C}_0/\{g(\widehat{F}_j)w_j\}]) = 1$. Since g and k are continuous and strictly increasing functions, it suffices to show that $\prod_{j=1}^{n} g(\widetilde{F}_{0j}) > 1$ and

$\prod_{j=1}^{n} g(\widetilde{F}_{\mathcal{C}j}) \leq 1$ for some $\mathcal{C}$. Obviously $\prod_{j=1}^{n} g(\widetilde{F}_{0j}) > 1$ since $g\{k^{-1}(0)\} > g(0) = 1$. Now let $\mathcal{C}_1 = k(0)\max_j\{g(\widehat{F}_j)w_j\} < 0$. Then for all $j, \mathcal{C}_1/\{g(\widehat{F}_j)w_j\} \leq k(0)$ and thus $g(k^{-1}[\mathcal{C}_1/\{g(\widehat{F}_j)w_j\}]) \leq g(0) = 1$. Then by continuity of g and k, there exists $\mathcal{C}_0 \in [\mathcal{C}_1, 0)$ such that $\prod_{j=1}^{n} g(\widetilde{F}_{\mathcal{C}_0 j}) = 1$. Thus, the constructed $\widetilde{\boldsymbol{F}}_{\mathcal{C}_0}$ possesses several properties: (i) $g(\widehat{F}_j)k(\widetilde{F}_{\mathcal{C}_0 j})w_j = \mathcal{C}_0$; (ii) $\prod_{j=1}^{n} g(\widetilde{F}_{\mathcal{C}_0 j}) = 1$; and (iii) $k(\widetilde{F}_{\mathcal{C}_0 j}) < 0$ and hence $\widetilde{\boldsymbol{F}}_{\mathcal{C}_0} \in \Omega$. It then follows from the AM-GM inequality that

$$\begin{aligned}\sum_{j=1}^{n} g(\widetilde{F}_{\mathcal{C}_0 j})\{-k(\widetilde{F}_{\mathcal{C}_0 j})\}w_j \geq n\left[\prod_{j=1}^{n} g(\widetilde{F}_{\mathcal{C}_0 j})\{-k(\widetilde{F}_{\mathcal{C}_0 j})\}w_j\right]^{n^{-1}} &= n\left[\prod_{j=1}^{n} g(\widehat{F}_j)\{-k(\widetilde{F}_{\mathcal{C}_0 j})\}w_j\right]^{n^{-1}} \\ &= -n\mathcal{C}_0,\end{aligned}$$

where we used the fact that $\prod_{j=1}^{n} g(\widetilde{F}_{\mathcal{C}_0 j}) = \prod_{j=1}^{n} g(\widehat{F}_j) = 1$. This, together with (A.0.4), implies that

$$n\mathcal{C}_0 \geq \sum_{j=1}^{n} g(\widetilde{F}_{\mathcal{C}_0 j})k(\widetilde{F}_{\mathcal{C}_0 j})w_j \geq \sum_{j=1}^{n} g(\widehat{F}_j)k(\widetilde{F}_{\mathcal{C}_0 j})w_j = n\mathcal{C}_0$$

and hence $n\mathcal{C}_0 = \sum_{j=1}^{n} g(\widetilde{F}_{\mathcal{C}_0 j})k(\widetilde{F}_{\mathcal{C}_0 j})w_j$. Thus, the equality holds in the AM-GM inequality above, which also implies that $g(\widetilde{F}_{\mathcal{C}_0 j})k(\widetilde{F}_{\mathcal{C}_0 j})w_j = \mathcal{C}_0$. Since $g(\widehat{F}_j)k(\widetilde{F}_{\mathcal{C}_0 j})w_j = \mathcal{C}_0$, $k(\widetilde{F}_{\mathcal{C}_0 j}) \neq 0$ and g is strictly increasing, we have $g(\widehat{F}_j) = g(\widetilde{F}_{\mathcal{C}_0 j})$ and hence $\widehat{F}_j = \widetilde{F}_{\mathcal{C}_0 j}$. Therefore,

$$g(\widehat{F}_j)k(\widehat{F}_j)w_j = H_\phi(\widehat{F}_j)w_j = \mathcal{C}_0.$$

Obviously if $H_\phi(\cdot)$ is strictly monotone then $\widehat{F}_j = H_\phi^{-1}(\mathcal{C}_0/w_j)$ which is unique. □

Proof of Proposition 2.2.2. The function ϕ is decreasing on the set S, as from (2.2.5) for any $x \leq$

$x', x, x' \in S$ we have:

$$\phi(x) - \phi(x') \geq (g(x) - g(x'))k(x') \geq 0.$$

Furthermore, it follows from Theorem 2.2.1, that $\widehat{F}_j \in S$ since $k(\widehat{F}_j) < 0$ for all j. Next we show that if $w_\jmath > w_j$ we must have $\phi(\widehat{F}_\jmath) \leq \phi(\widehat{F}_j)$. This observation follows since:

$$\phi(\widehat{F}_\jmath)w_\jmath + \phi(\widehat{F}_j)w_j \leq \phi(\widehat{F}_j)w_\jmath + \phi(\widehat{F}_\jmath)w_\jmath,$$

or else $\widehat{\boldsymbol{F}}$ cannot be a minimum of (2.2.8), as we can swap $\widehat{F}_\jmath$ and $\widehat{F}_j$ to obtain a strictly smaller value while still satisfying the constraint. Furthermore by Theorem 2.2.1, $w_\jmath \neq w_j$ implies that $\widehat{F}_\jmath \neq \widehat{F}_j$ because otherwise $H_\phi(\widehat{F}_\jmath) = H_\phi(\widehat{F}_j)$ and hence $w_\jmath = w_j$ by (2.2.9). Since ϕ is strictly decreasing on S it also implies $\phi(\widehat{F}_\jmath) \neq \phi(\widehat{F}_j)$. Hence $w_\jmath > w_j$ implies $\phi(\widehat{F}_\jmath) < \phi(\widehat{F}_j)$. Finally, the last observation gives:

$$\operatorname*{argmin}_{j\in\{1,\ldots,n\}} \phi(\widehat{F}_j) = \operatorname*{argmax}_{j\in\{1,\ldots,n\}} w_j.$$

The fact that ϕ is decreasing on S completes the proof. □

Proof of Theorem 2.2.3. To show that a finite minimizer $\widehat{\boldsymbol{F}}$ exists, it suffices to show that $g(\widehat{F}_j)$ is finite and bounded away from 0, for $j = 1, \ldots, n$. To this end, we note that the condition (2.2.11) is equivalent to,

$$\lim_{x\downarrow 0} c_1\phi(g^{-1}(x)) + c_2\phi(g^{-1}(x^{-(n-1)})) = +\infty \quad \text{for all } c_1, c_2 > 0. \tag{A.0.5}$$

We next show that at the minimizer $\widehat{\boldsymbol{F}}$, $\widehat{m} = \min_j g(\widehat{F}_j) = g(\widehat{F}_{j^*})$ is bounded away from 0, where $j^* = \operatorname{argmin}_j g(\widehat{F}_j)$. Since $1 = \prod_{j=1}^n g(\widehat{F}_j) \geq g(\widehat{F}_j)\widehat{m}^{n-1}$, we have $\widehat{F}_j \leq g^{-1}(\widehat{m}^{-(n-1)})$ for

$j = 1, ..., n$. If ϕ is *decreasing* over $\mathbb{R}$, then

$$\phi(0)\sum_{j=1}^{n} w_j \geq \sum_{j=1}^{n} \phi(\widehat{F}_j)w_j = w_{j^*}\phi\{g^{-1}(\widehat{m})\} + \sum_{j\neq j^*} w_j\phi(\widehat{F}_j)$$
$$\geq w_{j^*}\phi\{g^{-1}(\widehat{m})\} + \sum_{j\neq j^*} w_j\phi\{g^{-1}(\widehat{m}^{-(n-1)})\}.$$

From (A.0.5) with $c_1 = w_{j^*}$ and $c_2 = \sum_{j\neq j^*} w_j$, we conclude that $\widehat{m}$ must be bounded away from 0 since $\sum_{j=1}^{n} \phi(\widehat{F}_j)w_j \to \infty$ if $\widehat{m} \to 0$. Thus, there exists $m_0 > 0$ such that $\widehat{m} \geq m_0$ and consequently

$$0 < m_0 \leq g(\widehat{F}_j) \leq m_0^{-(n-1)} < \infty, \quad j = 1, ..., n.$$

Now, if ϕ is not decreasing on the whole $\mathbb{R}$, then there must exist $F^* < \infty$ such that $k(F^*) = 0$ since ϕ is strictly decreasing on $S = \{z : k(z) \leq 0\}$.

Now we show that $\widehat{\boldsymbol{F}} \in \Omega \equiv \{\boldsymbol{F} = (F_1, ..., F_n) : F_j \in S, j = 1, ..., n\}$ as defined in Theorem 2.2.1. To this end, we note that ϕ is strictly decreasing on S and $(-\infty, 0] \subset S$. We next argue by contradiction that $\widehat{F}_j \in S$ or equivalently $\widehat{F}_j \leq F^*$ for all j. For any $F > F^*$, $\phi(F) - \phi(F^*) \geq \{g(F) - g(F^*)\}k(F^*) = 0$ by (2.2.5). Let $\mathcal{A} = \{j : \widehat{F}_j > F^*\}$ and $\widehat{F}_j^* = I(j \in \mathcal{A})F^* + I(j \notin \mathcal{A})\widehat{F}_j$. If $\mathcal{A}$ is not an empty set, then $\sum_{j=1}^{n} \phi(\widehat{F}_j^*)w_j \leq \sum_{j=1}^{n} \phi(\widehat{F}_j)w_j$ and $\prod_{j=1}^{n} g(\widehat{F}_j^*) < 1$. Since $g(F^*) > 1$, there must exist some $\widehat{F}_j^{**}$ with $F^* \geq \widehat{F}_j^{**} \geq \widehat{F}_j$ for $j \notin \mathcal{A}$ and $\widehat{F}_j^{**} = F^*$ for $j \in \mathcal{A}$ such that $\prod_{j=1}^{n} g(\widehat{F}_j^{**}) = 1$ and $F^* \geq \widehat{F}_j^{**} > \widehat{F}_j$ for some $j \notin \mathcal{A}$. Since ϕ is strictly decreasing on S, $\sum_{j=1}^{n} \phi(\widehat{F}_j^{**})w_j < \sum_{j=1}^{n} \phi(\widehat{F}_j^*)w_j \leq \sum_{j=1}^{n} \phi(\widehat{F}_j)w_j$, which contradicts that $\widehat{\boldsymbol{F}}$ is the minimum. Therefore, $\widehat{\boldsymbol{F}} \in \Omega$.

Thus $\widehat{F}_j \leq F^*$ and $g(\widehat{F}_j) \leq g(F^*) = m_1 \in (0, \infty)$. On the other hand, since $\prod_{j=1}^{n} g(\widehat{F}_j) = 1$, we have $g(\widehat{F}_j) \geq m_1^{-(n-1)}$ and thus $g(\widehat{F}_j)$ is also bounded away from 0 and finite.

Remark A.0.2. *As a useful remark we mention that the same argument shows that given any finite*

vector $\widehat{\boldsymbol{F}}$, the vectors $\boldsymbol{F}$ with $\sum_j \phi(F_j)w_j \leq \sum_j \phi(\widehat{F}_j)w_j$ are located on a compact set (provided that $\widehat{F}_j < F^*$ for all j in the second case).

□

Lemma A.0.3. *Any loss function ϕ satisfying (2.2.5) with $g = \exp$, and either i. or ii. from Theorem 2.2.3 is classification calibrated in the two class case.*

Remark A.0.4. *Recall that a loss function ϕ is classification calibrated in the two class case if, for any point $w_1 + w_2 = 1$ with $w_1 \neq \frac{1}{2}$ and $w_1, w_2 > 0$, we have:*

$$\inf_{x \in \mathbb{R}} (w_1\phi(x) + w_2\phi(-x)) > \inf_{x : x(2w_1 - 1) \leq 0} (w_1\phi(x) + w_2\phi(-x)).$$

Proof of Lemma A.0.3. Denote the two (distinct) class probabilities with $w_1 + w_2 = 1$. Without loss of generality we distinguish two cases: $w_1 > w_2 > 0$ and $w_1 = 1, w_2 = 0$. First, consider the case when $w_1 > w_2 > 0$. Since the conditions of Theorem 2.2.3 hold, we know that the optimization problem (2.2.8) has a minimum, and hence by Proposition 2.2.2 we have that $\operatorname{argmax}_{j \in \{1,2\}} \widehat{F}_j \subseteq \{1\}$. Hence it follows that $\widehat{F}_1 > 0, \widehat{F}_2 < 0$ at the minimum. This implies that inequality in Remark A.0.4 is strict.

Next assume that $w_1 = 1, w_2 = 0$. This case is not covered by our results as we assume that the probabilities are bounded away from 0. As we argued earlier ϕ is strictly decreasing on the set S, where by assumption $(-\infty, 0] \subsetneq S$. Thus:

$$\widehat{\boldsymbol{F}} = \operatorname{argmin}_{\boldsymbol{F} : F_1 + F_2 = 0} w_1\phi(F_1) + w_2\phi(F_2) = \operatorname{argmin}_{\boldsymbol{F} : F_1 + F_2 = 0} \phi(F_1),$$

we must have $\widehat{F}_1 > 0$ and hence $\phi(0) > \phi(\widehat{F}_1)$. This finishes the proof. □

Lemma A.0.5. *Let $\boldsymbol{F}^{(m)}$ be defined as in iteration (2.3.1) starting from $\boldsymbol{F}^{(0)} = 0$. Then we must have $\boldsymbol{F}^{(m)} \in \Omega$ for all m, where $\Omega = \{\boldsymbol{F} = (F_1, \ldots, F_n) : F_j \in S, j = 1, ..., n\}$.*

Proof of Lemma A.o.5. We show the statement by induction. By definition $\boldsymbol{F}^{(0)} \in \Omega$. Assume that $\boldsymbol{F}^{(m-1)} \in \Omega$ for some $m \geq 1$. We now show that $\boldsymbol{F}^{(m)} \in \Omega$. To arrive at a contradiction, assume the contrary. Let $\mathcal{A} = \{j : F_j^{(m)} > F^*\} \neq \emptyset$. Define $F_j^{*(m)} = I(j \in \mathcal{A})F_j^{(m-1)} + I(j \notin \mathcal{A})F_j^{(m)}$. Since $\boldsymbol{F}^{(m-1)} \in \Omega$, it follows that $\boldsymbol{F}^{*(m)} \in \Omega$ and $\prod_{j=1}^n g(\boldsymbol{F}_j^{*(m)}) < 1$. More importantly, observe that for all $j \in \mathcal{A}$ we have:

$$0 = (g(F_j^{(m-1)}) - g(F_j^{*(m)}))k(F_j^{*(m)})w_j > (g(F_j^{(m-1)}) - g(F_j^{(m)}))k(F_j^{(m)})w_j,$$

as $g(F_j^{(m-1)}) \leq g(F^*) < g(F_j^{(m)})$ and $k(F_j^{(m)}) > k(F^*) = 0$, and hence:

$$\sum_{j=1}^n (g(F_j^{(m-1)}) - g(F_j^{*(m)}))k(F_j^{*(m)})w_j > \sum_{j=1}^n (g(F_j^{(m-1)}) - g(F_j^{(m)}))k(F_j^{(m)})w_j.$$

Define the index set $\mathcal{B} = \{j : F_j^{*(m)} < F_j^{(m-1)}\}$. Since $\prod_{j=1}^n g(\boldsymbol{F}_j^{*(m)}) < 1$ and $\boldsymbol{F}^{(m-1)} \in \Omega$ it follows that $\mathcal{B}$ is not empty and $\mathcal{A} \cap \mathcal{B} = \emptyset$. Next for $\lambda \in [0, 1]$ define for all j:

$$F_j^{*(m),\lambda} := [I(j \in \mathcal{A}) + I(j \notin \mathcal{A})I(j \notin \mathcal{B})]F_j^{*(m)} + I(j \in \mathcal{B})((1-\lambda)F_j^{*(m)} + \lambda F_j^{(m-1)}).$$

Note that when $\lambda = 0$, we have $F_j^{*(m),0} \equiv F_j^{*(m)}$. Now we show that for any $\lambda \in [0, 1]$ the following inequality holds:

$$\sum_{j=1}^n (g(F_j^{(m-1)}) - g(F_j^{*(m),\lambda}))k(F_j^{*(m),\lambda})w_j \geq \sum_{j=1}^n (g(F_j^{(m-1)}) - g(F_j^{*(m)}))k(F_j^{*(m)})w_j. \tag{A.o.6}$$

For any $\lambda \in (0, 1]$: $F_j^{*(m),\lambda} \neq F_j^{*(m),\lambda}$ iff $j \in \mathcal{B}$. Next note that for any j the function $(g(F_j^{(m-1)}) - g(x))k(x)w_j$ is an increasing function for $x \leq F_j^{(m-1)}$. The last two observations imply (A.o.6). Finally since $\prod_{j=1}^n g(F_j^{*(m),0}) = \prod_{j=1}^n g(F_j^{*(m)}) < 1$ and $\prod_{j=1}^n g(F_j^{*(m),1}) \geq$

$\prod_{j=1}^{n} g(F_j^{(m-1)}) = 1$, by the continuity of g there exists a $\lambda \in [0,1]$ such that $\prod_{j=1}^{n} g(F_j^{*(m),\lambda}) = 1$. These facts and inequality (A.0.6) imply that $\boldsymbol{F}^{(m)}$ would not be a maximum in the iteration which is a contradiction. □

Proof of Theorem 2.3.1. By construction we have that on the m^{th} iteration the value $\boldsymbol{F}^{(m)}$ satisfies $\prod_{j=1}^{m} g(F_j^{(m)}) = 1$, and Lemma A.0.5 guarantees that $\boldsymbol{F}^{(m)} \in \Omega$ for all m. Hence, since $F_j^{(m)}$ are viable values for $F_j^{(m+1)}$, the iteration also guarantees that:

$$\sum_{j=1}^{n} \{\phi(F_j^{(m)}) - \phi(F_j^{(m+1)})\} w_j \geq \sum_{j=1}^{n} \{g(F_j^{(m)}) - g(F_j^{(m+1)})\} k(F_j^{(m+1)}) w_j \geq 0.$$

Now, from Remark A.0.2, $F_j^{(m+1)}$ lie on a compact set for all j, since for our starting point we have $\boldsymbol{F}_j^{(0)} \equiv 0 \in \Omega$. Therefore there must exist a subsequence $\{m_\ell, \ell = 1, ...\}$ such that $\boldsymbol{F}^{(m_\ell)}$ converges coordinate-wise on this subsequence, and denote with $\boldsymbol{F}^*$ its limit.

The function ϕ is continuous and hence we have that $\sum_{j=1}^{n} \phi(F_j^{(m_\ell)}) w_j - \sum_{j=1}^{n} \phi(F_j^{(m_{\ell+1})}) w_j \to 0$. However by the construction of our iteration, the sequences $\sum_{j=1}^{n} \phi(F_j^{(m_{\ell+1})}) w_j$ are decreasing for all ℓ. Therefore we have that: $\sum_{j=1}^{n} \phi(F_j^{(m)}) w_j - \sum_{j=1}^{n} \phi(F_j^{(m+1)}) w_j \to 0$ holds for all m, not only on the subsequence. But this implies that $\sum_{j=1}^{n} (g(F_j^{(m)}) - g(F_j^{(m+1)})) k(F_j^{(m+1)}) w_j \to 0$, which again by the construction is non-negative for all m. Take m_ℓ in place of m in the limit above, and let L be the set of all limit points of $\boldsymbol{F}^{(m_\ell+1)}$. By our construction we have the following inequality holding for any point $\boldsymbol{F}^l \in L$:

$$0 = \sum_{j=1}^{n} \{g(F_j^*) - g(F_j^l)\} k(F_j^l) w_j \geq \sum_{j=1}^{n} \{g(F_j^*) - g(F_j)\} k(F_j) w_j, \tag{A.0.7}$$

for any $\boldsymbol{F} \in \Omega$ with $\prod_{j=1}^{n} g(F_j) = 1$. Just as in the proof of Theorem 2.2.1 select $\widetilde{\boldsymbol{F}}$ so that

$g(F_j^*)k(\widetilde{F}_j)w_j = \mathcal{C}$ for all j for some $\mathcal{C} < 0$. By the AM-GM inequality we get:

$$\sum_{j=1}^{n} g(\widetilde{F}_j)\{-k(\widetilde{F}_j)\}w_j \geq n\left[\prod_{j=1}^{n} g(\widetilde{F}_j)\{-k(\widetilde{F}_j)\}w_j\right]^{n^{-1}} = n\left[\prod_{j=1}^{n} g(F_j^*)\{-k(\widetilde{F}_j)\}w_j\right]^{n^{-1}}$$
$$= \sum_{j=1}^{n} g(F_j^*)\{-k(\widetilde{F}_j)\}w_j = -n\mathcal{C}.$$

Now by (A.0.7) it follows that equality must be achieved in the preceding display, which implies that $g(F_j^*)k(\widetilde{F}_j)w_j = \mathcal{C} = g(\widetilde{F}_j)k(\widetilde{F}_j)w_j$ and yields $\widetilde{F}_j = F_j^*$ for all j. Hence $g(F_j^*)k(F_j^*)w_j = \mathcal{C}$ for all j.

Thus we showed that on subsequences the iteration converges to points satisfying the equality described above. We are left to show, that all these subsequences converge to the same point. Next, take equation (A.0.7). By what we showed it follows that for any $\boldsymbol{F}^l \in L$, we have that $g(F_j^l)k(F_j^l)w_j = \mathcal{C}^l$ for some $\mathcal{C}^l < 0$. Then we have:

$$\sum_{j=1}^{n} g(F_j^*)\{-k(F_j^l)\}w_j \geq n\left[\prod_{j=1}^{n} g(F_j^*)\{-k(F_j^l)\}w_j\right]^{n^{-1}} = n\left[\prod_{j=1}^{n} g(F_j^l)\{-k(F_j^l)\}w_j\right]^{n^{-1}}$$
$$= \sum_{j=1}^{n} g(F_j^l)\{-k(F_j^l)\}w_j = -n\mathcal{C}^l.$$

Equation (A.0.7) implies that the above inequality is in fact equality which shows that:

$$g(F_j^*)k(F_j^l) = g(F_j^l)k(F_j^l) \quad \text{for all } j.$$

Thus since $k(F_j^l) \neq 0$ (recall that all values on the iteration $\boldsymbol{F}^{(m)} \in \Omega$) we conclude that $g(F_j^*) = g(F_j^l)$, and hence $\boldsymbol{F}^* = \boldsymbol{F}^l$. This shows that for any converging subsequence m_ℓ the limiting value coincides with that of the sequence $m_\ell + 1$, which finishes our proof.

□

Proof of Proposition 2.3.4. It is sufficient to show that for all $\boldsymbol{F} \in \mathcal{G}^*$ we have:

$$\sum_{i=1}^{N} \boldsymbol{Y}_{C_i}^{\intercal} \boldsymbol{F}(\boldsymbol{X}_i) \dot{\phi}(\boldsymbol{Y}_{C_i}^{\intercal} \boldsymbol{F}^{(\infty)}(\boldsymbol{X}_i)) \geq 0.$$

The condition above is sufficient, because of the looping closure of $\mathcal{G}$. Writing the inequality for all "looped" versions of $\boldsymbol{F}$, and noting that the sum up to 0, gives us that the inequality is in fact an equality.

Note that with each iteration (2.3.4), we decrease the value of the target function. This can be seen by the following inequality:

$$\sum_{i=1}^{N} \phi(\boldsymbol{Y}_{C_i}^{\intercal} \boldsymbol{F}^{(m-1)}(\boldsymbol{X}_i)) - \sum_{i=1}^{N} \phi(\boldsymbol{Y}_{C_i}^{\intercal} \boldsymbol{F}^{(m)}(\boldsymbol{X}_i)) \geq \sum_{i=1}^{N} [\exp(-\beta \boldsymbol{Y}_{C_i}^{\intercal} \boldsymbol{F}(\boldsymbol{X}_i)) - 1] \dot{\phi}(\boldsymbol{Y}_{C_i}^{\intercal} \boldsymbol{F}^{(m)}(\boldsymbol{X}_i)) \geq 0,$$

where $\boldsymbol{F}^{(m)} = \boldsymbol{F}^{(m-1)} + \beta \boldsymbol{F}$. As a remark, the inequality in the preceding display holds, since ϕ is decreasing, and thus by (2.2.5) we have $S = \mathbb{R}$.

Take a limiting point* $\boldsymbol{F}^{(\infty)}$ of iteration (2.3.4), where it is possible having coordinates of $\boldsymbol{F}^{(\infty)}(\boldsymbol{X}_i)$ equal to $\pm\infty$ for some i. Since ϕ is bounded from below, by our previous observation we have that for any $\beta \geq 0$ and $\boldsymbol{F} \in \mathcal{G}^*$:

$$\sum_{i=1}^{N} \phi(\boldsymbol{Y}_{C_i}^{\intercal} \boldsymbol{F}^{(\infty)}(\boldsymbol{X}_i)) - \sum_{i=1}^{N} \phi(\boldsymbol{Y}_{C_i}^{\intercal} \boldsymbol{F}^{(\infty)}(\boldsymbol{X}_i) + \beta \boldsymbol{Y}_{C_i}^{\intercal} \boldsymbol{F}(\boldsymbol{X}_i)) \leq 0.$$

*The existence of a limiting point is guaranteed as any sequence contains a monotone subsequence.

Let $\mathcal{A} = \{i : |\boldsymbol{Y}_{C_i}^\intercal \boldsymbol{F}^{(\infty)}(\boldsymbol{X}_i)| \neq \infty\}$. Then the latter inequality also implies that:

$$\sum_{i\in\mathcal{A}} \phi(\boldsymbol{Y}_{C_i}^\intercal \boldsymbol{F}^{(\infty)}(\boldsymbol{X}_i)) - \sum_{i\in\mathcal{A}} \phi(\boldsymbol{Y}_{C_i}^\intercal \boldsymbol{F}^{(\infty)}(\boldsymbol{X}_i) + \beta \boldsymbol{Y}_{C_i}^\intercal \boldsymbol{F}(\boldsymbol{X}_i)) \leq 0.^\dagger$$

Applying inequality (2.2.5) the above implies:

$$\sum_{i\in\mathcal{A}} [\exp(-\beta \boldsymbol{Y}_{C_i}^\intercal \boldsymbol{F}(\boldsymbol{X}_i)) - 1]\dot{\phi}(\boldsymbol{Y}_{C_i}^\intercal \boldsymbol{F}^{(\infty)}(\boldsymbol{X}_i) + \beta \boldsymbol{Y}_{C_i}^\intercal \boldsymbol{F}(\boldsymbol{X}_i)) \leq 0,$$

and after a Taylor expansion of the exponent, and division by $\beta \geq 0$ we get:

$$\begin{aligned}
&\sum_{i\in\mathcal{A}} -\boldsymbol{Y}_{C_i}^\intercal \boldsymbol{F}(\boldsymbol{X}_i)\dot{\phi}(\boldsymbol{Y}_{C_i}^\intercal \boldsymbol{F}^{(\infty)}(\boldsymbol{X}_i)) \\
&+\sum_{i\in\mathcal{A}} -\boldsymbol{Y}_{C_i}^\intercal \boldsymbol{F}(\boldsymbol{X}_i)[\dot{\phi}(\boldsymbol{Y}_{C_i}^\intercal \boldsymbol{F}^{(\infty)}(\boldsymbol{X}_i) + \beta \boldsymbol{Y}_{C_i}^\intercal \boldsymbol{F}(\boldsymbol{X}_i)) - \dot{\phi}(\boldsymbol{Y}_{C_i}^\intercal \boldsymbol{F}^{(\infty)}(\boldsymbol{X}_i))] \\
&+O(\beta)\sum_{i\in\mathcal{A}} \dot{\phi}(\boldsymbol{Y}_{C_i}^\intercal \boldsymbol{F}^{(\infty)}(\boldsymbol{X}_i) + \beta \boldsymbol{Y}_{C_i}^\intercal \boldsymbol{F}(\boldsymbol{X}_i)) \leq 0.
\end{aligned}$$

Letting $\beta \to 0$, by the continuity of $\dot{\phi}$ we get:

$$\sum_{i\in\mathcal{A}} \boldsymbol{Y}_{C_i}^\intercal \boldsymbol{F}(\boldsymbol{X}_i)\dot{\phi}(\boldsymbol{Y}_{C_i}^\intercal \boldsymbol{F}^{(\infty)}(\boldsymbol{X}_i)) \geq 0. \tag{A.o.8}$$

Next we argue that $\dot{\phi}(+\infty) = \lim_{x\to+\infty} \dot{\phi}(x) = 0$. As stated in the main text $\dot{\phi}(x) = k(x)e^x$. Let $K = \inf_{x\in\mathbb{R}} \phi(x)$. For any $\varepsilon > 0$, take any point x' so that $\phi(x') - K \leq \varepsilon$. Then for any $x \in \mathbb{R}$, by (2.2.5):

$$\varepsilon \geq \phi(x') - \phi(x) \geq (e^{x'} - e^x)k(x),$$

and thus $\varepsilon - e^{x'}k(x) \geq -\dot{\phi}(x) \geq 0$. Taking the limit $x \to +\infty$ and letting $\varepsilon \to 0$ shows that

†Observe that since ϕ is bounded from below the values of ϕ at singular points i.e. $\phi(\pm\infty)$ have to be bounded.

$\dot{\phi}(+\infty) = 0$.

Now consider two cases for ϕ. Suppose that ϕ is unbounded from above. We argue that $\boldsymbol{Y}_{C_i}^\intercal \boldsymbol{F}^{(\infty)}(\boldsymbol{X}_i) \neq -\infty$ for all i. Since we start form the point 0, and as we argued we are decreasing the target function we have that:

$$N\phi(0) \geq \sum_{i=1}^{N} \phi(\boldsymbol{Y}_{C_i}^\intercal \boldsymbol{F}^{(\infty)}(\boldsymbol{X}_i)) \geq \max_i \phi(\boldsymbol{Y}_{C_i}^\intercal \boldsymbol{F}^{(\infty)}(\boldsymbol{X}_i)) + (N-1)K,$$

and hence $\max_i \phi(\boldsymbol{Y}_{C_i}^\intercal \boldsymbol{F}^{(\infty)}(\boldsymbol{X}_i)) \leq N\phi(0) - (N-1)K$. Since ϕ is decreasing and unbounded from above it follows that $\boldsymbol{Y}_{C_i}^\intercal \boldsymbol{F}^{(\infty)}(\boldsymbol{X}_i) \neq -\infty$ for all i. In the second case suppose that ϕ is bounded from above, and let $M = \sup_{x\in\mathbb{R}} \phi(x)$. We show that $\dot{\phi}(-\infty) = 0$. For any $\varepsilon > 0$ take x so that $\varepsilon \geq M - \phi(x)$. Applying (2.2.5) for any $x' \in \mathbb{R}$ yields:

$$\varepsilon \geq \phi(x') - \phi(x) \geq (e^{x'} - e^x)k(x).$$

This gives $\varepsilon - e^{x'}k(x) \geq -\dot{\phi}(x) \geq 0$. Taking $x' \to -\infty$ gives that $\varepsilon \geq -\dot{\phi}(x) \geq 0$ for any x such that $\varepsilon \geq M - \phi(x)$. Since ϕ is decreasing we are allowed to take the limit $x \to -\infty$, and taking $\varepsilon \to 0$ shows that $\dot{\phi}(-\infty) = 0$. In any case, all of the above arguments imply that we can substitute $\mathcal{A}$ in (A.0.8) to the whole index set $\{1, \ldots, N\}$, to finally conclude:

$$\sum_{i=1}^{N} \boldsymbol{Y}_{C_i}^\intercal \boldsymbol{F}(\boldsymbol{X}_i)\dot{\phi}(\boldsymbol{Y}_{C_i}^\intercal \boldsymbol{F}^{(\infty)}(\boldsymbol{X}_i)) \geq 0,$$

for all $\boldsymbol{F} \in \mathcal{G}^*$. As argued in the beginning the looping closure gives us that in fact the "$\geq$" can be replaced with "$=$". This concludes the proof.

□

Proof of Proposition 2.3.5. First consider the case when $I = N$. Denote with $e_1^1, e_1^2, \ldots, e_1^N$

the positive coordinates of $\boldsymbol{e}_1$. For any vector $\boldsymbol{v}$ which is a solution to $\boldsymbol{D}^\intercal\boldsymbol{\alpha} = \boldsymbol{v}$ we must have $\sum_{i=1}^N \boldsymbol{e}_1^i v_i = 0$. Clearly then, if $\boldsymbol{v}$ is non-zero then some of the v_i need to be negative. Let $l = \min v_i$. We know that $l < 0$. This immediately implies an upper bound on the maximal positive v_i — $\max_i v_i \leq |l|[\sum_{i=1}^N e_1^i / \min_j e_1^j - 1]$. We now show that $|l|$ is bounded, for all vectors $\boldsymbol{v}$ such that $\sum_{i=1}^N \phi(v_i) \leq N\phi(0)$. Note that $\sum_{i=1}^N \phi(v_i) \geq \phi(l) + (N-1)\phi(|l|[\sum_{i=1}^N e_1^i / \min_j e_1^j - 1])$, and thus (2.3.8) gives that $|l|$ is bounded. This in turn shows that all v_i are bounded as well.

We next consider the case where $I < N$. Without loss of generality, upon rearrangement, we can assume that the positive coordinates of $\boldsymbol{e}_1$ are located at the first I places. Let $\boldsymbol{E}_{N\times s} = (\boldsymbol{e}_1, \ldots, \boldsymbol{e}_s)$. Let $\widetilde{\boldsymbol{E}}_{(N-I)\times(s-1)}$ denote the sub-matrix corresponding to the 0 entries of $\boldsymbol{e}_1$ (excluding $\boldsymbol{e}_1$) (see (A.0.9) for a visualization). Note that the matrix $\widetilde{\boldsymbol{E}}$ cannot be of full column rank, because otherwise we would have that a vector with positive coordinates is inside the column space which is a contradiction (we can always scale it by a small number and add to $\boldsymbol{e}_1$). Thus we can eliminate all extra columns that do not contribute to the rank of $\widetilde{\boldsymbol{E}}$, by doing a linear manipulation on the columns of the whole matrix $\boldsymbol{E}$ (see (A.0.9)). In doing so, we can eliminate extra columns of the matrix $\widetilde{\boldsymbol{E}}$ so that we end up with a $\widetilde{\boldsymbol{E}}$ matrix where the number of non-zero columns matches the rank, and some columns of $\boldsymbol{E}$ have 0 coefficients on the lower part. Here, observe that the columns of $\boldsymbol{E}$ with 0 sub-columns in $\widetilde{\boldsymbol{E}}$, are part of the space $\mathrm{row}(\boldsymbol{D}_-)^\perp$, where $\boldsymbol{D}_-$ corresponds to the matrix $\boldsymbol{D}$ with observations corresponding to 0's of $\boldsymbol{e}_1$ removed.

We next note that if we discard the observations corresponding to 0 coordinates in e_1, and optimize the problem on the rest of the observations we will obtain some optimal solution $\boldsymbol{v} = (\widehat{v}_1, \ldots, \widehat{v}_I)^\intercal$, the entries of which are bounded as argued in the first case. We next show that we can populate the vector $\boldsymbol{v}$ with positive numbers $p_1, \ldots, p_{N-I}$ to $\boldsymbol{\nu} = (\widehat{v}_1, \ldots, \widehat{v}_I, p_1, \ldots, p_{N-I})^\intercal$, so that $\boldsymbol{\nu}$ is "perpendicular" to the matrix $\boldsymbol{E}$ (i.e. $\boldsymbol{E}^\intercal\boldsymbol{\nu} = 0$), and thus can be written in the form $\boldsymbol{D}^\intercal\boldsymbol{\alpha}$. Moreover, we will show that $p_1, \ldots, p_{N-I}$, can become arbitrarily large, which will com-

plete the proof.

$$\boldsymbol{E} = \begin{array}{c|c|c} & \boldsymbol{e}_1 & \boldsymbol{e}_2 \ \ldots \ \boldsymbol{e}_s \\ \hline I & e_1^1 \\ & \vdots \\ & e_1^I \\ \hline N-I & 0 \\ & \vdots \quad \widetilde{\boldsymbol{E}} \\ & 0 \end{array} \quad \rightarrow \quad \begin{array}{c|c|c|c} \boldsymbol{e}_1 & \ldots \ \widetilde{\boldsymbol{e}}_{l+1} & \widetilde{\boldsymbol{e}}_{l+2} & \ldots \ \widetilde{\boldsymbol{e}}_s \\ \hline e_1^1 & & & \\ \vdots & \boldsymbol{G} & & \\ e_1^I & & & \\ \hline 0 & & 0 \ \ldots \ 0 & \\ \vdots & \widetilde{\widetilde{\boldsymbol{E}}} & \vdots \ \ldots \ \vdots & \\ 0 & & 0 \ \ldots \ 0 & \end{array} \tag{A.0.9}$$

Note that the only part of the matrix $\boldsymbol{E}$ that would be potentially non-zero upon multiplication by $\boldsymbol{\nu}$ would be the part corresponding to the non-zero parts of $\widetilde{\boldsymbol{E}}$, because as we argued earlier the columns of $\boldsymbol{E}$ with 0 sub-columns in $\widetilde{\boldsymbol{E}}$ belong to $\text{row}(\boldsymbol{D}_-)^\perp$ and on the other hand $\boldsymbol{v} \in \text{row}(\boldsymbol{D}_-)$. Denote with $\widetilde{\widetilde{\boldsymbol{E}}}_{(N-I)\times l}$ the full-rank sub-matrix of $\widetilde{\boldsymbol{E}}$, where l is the rank of $\widetilde{\boldsymbol{E}}$, and let $\boldsymbol{G}_{I\times l}$ be the sub-matrix of $\boldsymbol{E}$ above $\widetilde{\widetilde{\boldsymbol{E}}}$ (see (A.0.9)). Clearly $l < N - I$ as otherwise there is a positive vector in the column space, and we argued previously that would be a contradiction with the maximality property of $\boldsymbol{e}_1$. We need to find a positive vector $\boldsymbol{p}$ such that $(\widetilde{\widetilde{\boldsymbol{E}}}_{(N-I)\times l})^\intercal \boldsymbol{p}_{(N-I)\times 1} = -(\boldsymbol{G}_{I\times l})^\intercal \boldsymbol{v}_{I\times 1} = \boldsymbol{K}_{l\times 1}$. Therefore the proof will be completed, if we can find arbitrary large positive vectors $\boldsymbol{p}$ solving the system $\widetilde{\widetilde{\boldsymbol{E}}}^\intercal \boldsymbol{p} = \boldsymbol{K}$, where $l < N - I$ and $\widetilde{\widetilde{\boldsymbol{E}}}^\intercal$ has the property that any non-zero linear combination of its rows results into a vector with at least one positive and one negative entry.

Since $l < N - I$, the linear system $\widetilde{\widetilde{\boldsymbol{E}}}^\intercal \boldsymbol{p} = \boldsymbol{K}$ has a solution. Consider the homogeneous system $\widetilde{\widetilde{\boldsymbol{E}}}^\intercal \boldsymbol{p} = 0$. We will show that the homogeneous equation admits arbitrary large positive solutions, which would complete the proof. Fix the value of the i^{th} parameter to be 1. The system then becomes $\widetilde{\widetilde{\boldsymbol{E}}}^\intercal_{-i} \boldsymbol{p}_{-i} = -\widetilde{\widetilde{\boldsymbol{e}}}_i$, where by indexing with $-i$ we mean removing the i^{th} column or element

and $\widetilde{\widetilde{e}}_i$ is the i^{th} column of $\widetilde{\widetilde{E}}^{\mathsf{T}}$. Next we apply Farkas's lemma to show that the last equation has a non-negative solution. Assume that there is a vector $\mathbf{y}_{l\times 1}$ such that $\widetilde{\widetilde{E}}_{-i}\mathbf{y} \geq 0$ (coordinate-wise) and $-\widetilde{\widetilde{e}}_i^{\mathsf{T}}\mathbf{y} < 0$. This is clearly a violation with the property that $\widetilde{\widetilde{E}}$ satisfies. Therefore by Farkas's lemma the equation $\widetilde{\widetilde{E}}_{-i}^{\mathsf{T}}\boldsymbol{p}_{-i} = -\widetilde{\widetilde{e}}_i$ has a non-negative solution. Since we can achieve this for any index i, averaging these solutions yields a positive solution to the homogeneous system $\widetilde{\widetilde{E}}^{\mathsf{T}}\boldsymbol{p} = 0$, and thus this system admits arbitrarily large positive solutions.

□

Proof of Theorem 2.3.6. Without loss of generality for the purposes of the proof we will consider $\mathcal{C}_+ = 1$ and $\mathcal{C}_- = -1/(n-1)$ (it's equivalent to rescaling the β in the iteration).

By the iteration's construction we know:

$$\varepsilon_m - \varepsilon_{m+1} \geq \max_{\beta\geq 0, \boldsymbol{F}\in\mathcal{G}^*} \sum_{i=1}^{N} \{e^{-\beta \boldsymbol{Y}_{C_i}^{\mathsf{T}}\boldsymbol{F}(\boldsymbol{X}_i)} - 1\}\dot{\phi}\{\boldsymbol{Y}_{C_i}^{\mathsf{T}}\boldsymbol{F}^{(m)}(\boldsymbol{X}_i) + \beta \boldsymbol{Y}_{C_i}^{\mathsf{T}}\boldsymbol{F}(\boldsymbol{X}_i)\}.$$

Note that we have the following simple inequality holding for $\exp(-x) \leq 1 - x + x^2$ for values of $-1/2 \leq x \leq 1/2$. Since $|\boldsymbol{Y}_{C_i}^{\mathsf{T}}\boldsymbol{F}(\boldsymbol{X}_i)| \leq \mathcal{C}_+ = 1$ and ϕ is decreasing, for values of $0 \leq \beta \leq 1/2$ we have that:

$$\begin{aligned}
&\sum_{i=1}^{N} \{e^{-\beta \boldsymbol{Y}_{C_i}^{\mathsf{T}}\boldsymbol{F}(\boldsymbol{X}_i)} - 1\}\dot{\phi}\{\boldsymbol{Y}_{C_i}^{\mathsf{T}}\boldsymbol{F}^{(m)}(\boldsymbol{X}_i) + \beta \boldsymbol{Y}_{C_i}^{\mathsf{T}}\boldsymbol{F}(\boldsymbol{X}_i)\} \\
&\geq -\sum_{i=1}^{N} (\beta \boldsymbol{Y}_{C_i}^{\mathsf{T}}\boldsymbol{F}(\boldsymbol{X}_i) - \beta^2)\dot{\phi}\{\boldsymbol{Y}_{C_i}^{\mathsf{T}}\boldsymbol{F}^{(m)}(\boldsymbol{X}_i) + \beta \boldsymbol{Y}_{C_i}^{\mathsf{T}}\boldsymbol{F}(\boldsymbol{X}_i)\}.
\end{aligned}$$

Let L denote the Lipschitz constant of $\dot{\phi}$ on the set $\mathcal{S}$. Consequently we have:

$$
\begin{aligned}
&-\sum_{i=1}^{N}(\beta \boldsymbol{Y}_{C_i}^{\intercal}\boldsymbol{F}(\boldsymbol{X}_i)-\beta^2)\dot{\phi}\{\boldsymbol{Y}_{C_i}^{\intercal}\boldsymbol{F}^{(m)}(\boldsymbol{X}_i)+\beta \boldsymbol{Y}_{C_i}^{\intercal}\boldsymbol{F}(\boldsymbol{X}_i)\}\\
&=\sum_{i=1}^{N}-(\beta \boldsymbol{Y}_{C_i}^{\intercal}\boldsymbol{F}(\boldsymbol{X}_i)-\beta^2)\dot{\phi}\{\boldsymbol{Y}_{C_i}^{\intercal}\boldsymbol{F}^{(m)}(\boldsymbol{X}_i)\}\\
&\quad-\sum_{i=1}^{N}(\beta \boldsymbol{Y}_{C_i}^{\intercal}\boldsymbol{F}(\boldsymbol{X}_i)-\beta^2)[\dot{\phi}\{\boldsymbol{Y}_{C_i}^{\intercal}\boldsymbol{F}^{(m)}(\boldsymbol{X}_i)+\beta \boldsymbol{Y}_{C_i}^{\intercal}\boldsymbol{F}(\boldsymbol{X}_i)\}-\dot{\phi}\{\boldsymbol{Y}_{C_i}^{\intercal}\boldsymbol{F}^{(m)}(\boldsymbol{X}_i)\}]\geq\\
&\quad\sum_{i=1}^{N}-(\beta \boldsymbol{Y}_{C_i}^{\intercal}\boldsymbol{F}(\boldsymbol{X}_i)-\beta^2)\dot{\phi}\{\boldsymbol{Y}_{C_i}^{\intercal}\boldsymbol{F}^{(m)}(\boldsymbol{X}_i)\}-L\left|\beta \boldsymbol{Y}_{C_i}^{\intercal}\boldsymbol{F}(\boldsymbol{X}_i)-\beta^2\right|\left|\beta \boldsymbol{Y}_{C_i}^{\intercal}\boldsymbol{F}(\boldsymbol{X}_i)\right|\geq\\
&\quad\sum_{i=1}^{N}-(\beta \boldsymbol{Y}_{C_i}^{\intercal}\boldsymbol{F}(\boldsymbol{X}_i)-\beta^2)\dot{\phi}\{\boldsymbol{Y}_{C_i}^{\intercal}\boldsymbol{F}^{(m)}(\boldsymbol{X}_i)\}-\frac{3}{2}LN\beta^2.
\end{aligned}
$$

Thus we have established:

$$
\varepsilon_m-\varepsilon_{m+1}\geq\beta\left(\sum_{i=1}^{N}(-\boldsymbol{Y}_{C_i}^{\intercal}\boldsymbol{F}(\boldsymbol{X}_i)+\beta)\dot{\phi}\{\boldsymbol{Y}_{C_i}^{\intercal}\boldsymbol{F}^{(m)}(\boldsymbol{X}_i)\}-\frac{3}{2}LN\beta\right),
$$

for any $0\leq\beta\leq 1/2$. We select β so that we maximize the RHS in the expression above. It turns out that this happens for:

$$
\beta_0=\frac{\frac{1}{2}\sum_{i=1}^{N}\boldsymbol{Y}_{C_i}^{\intercal}\boldsymbol{F}(\boldsymbol{X}_i)\dot{\phi}\{\boldsymbol{Y}_{C_i}^{\intercal}\boldsymbol{F}^{(m)}(\boldsymbol{X}_i)\}}{-\frac{3}{2}LN+\sum_{i=1}^{N}\dot{\phi}\{\boldsymbol{Y}_{C_i}^{\intercal}\boldsymbol{F}^{(m)}(\boldsymbol{X}_i)\}}.
$$

Since $\dot{\phi}$ is always negative and as we mentioned $\left|\boldsymbol{Y}_{C_i}^{\intercal}\boldsymbol{F}(\boldsymbol{X}_i)\right|\leq 1$, provided that the numerator is ≤ 0, we have that $0\leq\beta_0\leq 1/2$.

Then we would have:

$$\varepsilon_m - \varepsilon_{m+1} \geq -\frac{1}{2}\beta_0 \sum_{i=1}^{N} \boldsymbol{Y}_{C_i}^{\intercal} \boldsymbol{F}(\boldsymbol{X}_i) \dot{\phi}\{\boldsymbol{Y}_{C_i}^{\intercal} \boldsymbol{F}^{(m)}(\boldsymbol{X}_i)\}. \tag{A.0.10}$$

Next we show that there exists a classifier, such that the above expression is strictly positive, which will also ensure that $0 \leq \beta_0 \leq 1/2$ is in the correct range. Denote with B the total number of classifiers in the bag. Consider the representation $\boldsymbol{F}^*(\cdot) - \boldsymbol{F}^{(m)}(\cdot) = \sum_{j=1}^{B} \alpha_j \boldsymbol{F}_j(\cdot)$. Here the $\boldsymbol{\alpha}$ vector is any vector that yields a correct representation (note that we will have many possible $\boldsymbol{\alpha}$ vectors, in the case when $B > N$).

By convexity of ϕ we have:

$$\begin{aligned} -\varepsilon_m = \sum_{i=1}^{N} \phi(\boldsymbol{Y}_{C_i}^{\intercal} \boldsymbol{F}^*(\boldsymbol{X}_i)) - \phi(\boldsymbol{Y}_{C_i}^{\intercal} \boldsymbol{F}^{(m)}(\boldsymbol{X}_i)) &\geq \sum_{i=1}^{N} [\boldsymbol{Y}_{C_i}^{\intercal} \boldsymbol{F}^*(\boldsymbol{X}_i) - \boldsymbol{Y}_{C_i}^{\intercal} \boldsymbol{F}^{(m)}(\boldsymbol{X}_i)] \dot{\phi}(\boldsymbol{Y}_{C_i}^{\intercal} \boldsymbol{F}^{(m)}(\boldsymbol{X}_i)) \\ &= \sum_{j=1}^{B} \sum_{i=1}^{N} \alpha_j \boldsymbol{Y}_{C_i}^{\intercal} \boldsymbol{F}_j(\boldsymbol{X}_i) \dot{\phi}(\boldsymbol{Y}_{C_i}^{\intercal} \boldsymbol{F}^{(m)}(\boldsymbol{X}_i)). \end{aligned}$$

By the pigeonhole principle it is clear that there exists an index $j \in \{1, \ldots, B\}$ such that:

$$\frac{\varepsilon_m}{B \max_j |\alpha_j|} \leq \frac{\varepsilon_m}{B|\alpha_j|} \leq -\operatorname{sign}(\alpha_j) \sum_{i=1}^{N} \boldsymbol{Y}_{C_i}^{\intercal} \boldsymbol{F}_j(\boldsymbol{X}_i) \dot{\phi}(\boldsymbol{Y}_{C_i}^{\intercal} \boldsymbol{F}^{(m)}(\boldsymbol{X}_i)).$$

Now if $\operatorname{sign}(\alpha_j) = 1$ we already have a "decent" lower bound. Otherwise if $\operatorname{sign}(\alpha_j) = -1$, using the fact that the loop closed classifiers wrt to $\boldsymbol{F}_j$ sum up to o, we can claim that for one of the looped classifiers $\boldsymbol{F}_j^l$ we would have a bound:

$$\frac{\varepsilon_m}{B(n-1) \max_j |\alpha_j|} \leq -\sum_{i=1}^{N} \boldsymbol{Y}_{C_i}^{\intercal} \boldsymbol{F}_j^l(\boldsymbol{X}_i) \dot{\phi}(\boldsymbol{Y}_{C_i}^{\intercal} \boldsymbol{F}^{(m)}(\boldsymbol{X}_i)).$$

So that in both cases we established the existence of a classifier such that $F \in \mathcal{G}^*$ and:

$$\frac{\varepsilon_m}{B(n-1)\max_j |\alpha_j|} \leq -\sum_{i=1}^{N} \boldsymbol{Y}_{C_i}^\intercal \boldsymbol{F}(\boldsymbol{X}_i)\dot{\phi}(\boldsymbol{Y}_{C_i}^\intercal \boldsymbol{F}^{(m)}(\boldsymbol{X}_i))$$

We then know from (A.0.10) that:

$$\begin{aligned}\varepsilon_m - \varepsilon_{m+1} &\geq -\frac{1}{2}\beta_0 \sum_{i=1}^{N} \boldsymbol{Y}_{C_i}^\intercal \boldsymbol{F}(\boldsymbol{X}_i)\dot{\phi}\{\boldsymbol{Y}_{C_i}^\intercal \boldsymbol{F}^{(m)}(\boldsymbol{X}_i)\} \\ &\geq \frac{1}{4}\frac{\varepsilon_m^2}{B^2(n-1)^2 \max_j \alpha_j^2 (\frac{3}{2}LN - \sum_{i=1}^{N} \dot{\phi}\{\boldsymbol{Y}_{C_i}^\intercal \boldsymbol{F}^{(m)}(\boldsymbol{X}_i)\})}.\end{aligned}$$

Notice that the derivative is bounded on the set $\mathcal{S}$ and therefore collapsing all constants above into one constant say T we get the following:

$$\varepsilon_m - \varepsilon_{m+1} \geq \frac{\varepsilon_m^2}{T \max_j \alpha_j^2}.$$

Here T depends on the number of classifiers, number of classes, and the bound on the first derivative $\dot{\phi}$ on the set $\mathcal{S}$. We will proceed to bound the $\max_j \alpha_j^2$ for some of the representations from above.

Because on the set $\mathcal{S}$, ϕ is also strongly convex (with a constant say l), we have the following:

$$\begin{aligned}\varepsilon_m &= \sum_{i=1}^{N} \phi(\boldsymbol{Y}_{C_i}^\intercal \boldsymbol{F}^{(m)}(\boldsymbol{X}_i)) - \sum_{i=1}^{N} \phi(\boldsymbol{Y}_{C_i}^\intercal \boldsymbol{F}^{*}(\boldsymbol{X}_i)) \\ &\geq \sum_{i=1}^{N} [\boldsymbol{Y}_{C_i}^\intercal \boldsymbol{F}^{(m)}(\boldsymbol{X}_i) - \boldsymbol{Y}_{C_i}^\intercal \boldsymbol{F}^{*}(\boldsymbol{X}_i)]\dot{\phi}(\boldsymbol{Y}_{C_i}^\intercal \boldsymbol{F}^{*}(\boldsymbol{X}_i)) \\ &+ l \sum_{i=1}^{N} \left(\sum_{j=1}^{B} \alpha_j \boldsymbol{Y}_{C_i}^\intercal \boldsymbol{F}_j(\boldsymbol{X}_i) \right)^2.\end{aligned}$$

The expression $\sum_{i=1}^{N}[\boldsymbol{Y}_{C_i}^{\intercal}\boldsymbol{F}^{(m)}(\boldsymbol{X}_i) - \boldsymbol{Y}_{C_i}^{\intercal}\boldsymbol{F}^{*}(\boldsymbol{X}_i)]\dot{\phi}(\boldsymbol{Y}_{C_i}^{\intercal}\boldsymbol{F}^{*}(X_i))$ is 0, as $\boldsymbol{F}^{*}$ is the minimum, ϕ is convex and the classifier bag is closed under looping. Let $\boldsymbol{D} = \{\boldsymbol{Y}_{C_i}^{\intercal}\boldsymbol{F}_j(\boldsymbol{X}_i)\}_{j,i}$ is the $B \times N$ matrix, each entry of which is either $\mathcal{C}_+$ or $\mathcal{C}_-$. Let the rank of D is $r \leq \min(N, B)$. We then have $\sum_{i=1}^{N}\left(\sum_{j=1}^{B}\alpha_j\boldsymbol{Y}_{C_i}^{\intercal}\boldsymbol{F}_j(\boldsymbol{X}_i)\right)^2 = \boldsymbol{\alpha}^{\intercal}\boldsymbol{D}\boldsymbol{D}^{\intercal}\boldsymbol{\alpha}$. Since, all the bounds above are true for any of the $\boldsymbol{\alpha}$ representations, we could have picked the representation corresponding to the $r \times N$ sub matrix of $\boldsymbol{D}$, $\boldsymbol{D}_r$ with rank r for which the smallest eigenvalue of $\boldsymbol{D}_r\boldsymbol{D}_r^{\intercal}$ is the largest. Let this eigenvalue be $\lambda_r > 0$. For this eigenvalue and this choice of $\boldsymbol{\alpha}$ we clearly have $\boldsymbol{\alpha}^{\intercal}\boldsymbol{D}\boldsymbol{D}^{\intercal}\boldsymbol{\alpha} = \boldsymbol{\alpha}^{\intercal}\boldsymbol{D}_r\boldsymbol{D}_r^{\intercal}\boldsymbol{\alpha} \geq \lambda_r||\boldsymbol{\alpha}||_2^2 \geq \lambda_r \max_j \alpha_j^2$. (in the second equality we abuse notation deleting zeros from the $\boldsymbol{\alpha}$). Consequently we get:

$$\varepsilon_m \geq l\lambda_r \max_j \alpha_j^2.$$

Thus:

$$\begin{aligned}\varepsilon_{m+1} &\leq \varepsilon_m - \frac{\varepsilon_m^2}{T\max_j \alpha_j^2} \\ &\leq \varepsilon_m\left(1 - \frac{l\lambda_r}{T}\right).\end{aligned}$$

Since both $\varepsilon_{m+1}, \varepsilon_m \geq 0$ we must have $1 - \frac{l\lambda_r}{T} \geq 0$. Furthermore, by construction we have $\frac{l\lambda_r}{T} > 0$, which of course concludes the proof of the geometric rate.

□

Remark A.o.6. *It can be seen that even if we only assume that $\dot{\phi}$ is Lipschitz on $\mathcal{S}$ without the first derivative being bounded we can still obtain a geometric rate of convergence.*

B

Proofs for Chapter 3

B.1 SIR related proofs

Proof of Lemma 3.2.10. Here we deal with the expression, given in the sliced stability example. We start with the following observation:

$$\Phi^{-1}\left(\frac{1}{2}+\frac{q}{2}\right) \geq \frac{q}{2}\frac{1}{\phi(\Phi^{-1}\left(\frac{1}{2}+\frac{q}{2}\right))}+\frac{1}{2}\left(\frac{q}{2}\right)^2\frac{\Phi^{-1}\left(\frac{1}{2}+\frac{q}{2}\right)}{\phi^2\left(\Phi^{-1}\left(\frac{1}{2}+\frac{q}{2}\right)\right)}.$$

Which follows after an application of the mean value theorem, and noting that $\frac{d^3}{dx^3}\Phi^{-1}(x) = \frac{1+2(\Phi^{-1}(x))^2}{\phi^3(\Phi^{-1}(x))} \geq 0$. Thus for $q \in [0, 2\Phi(r) - 1)$ for some $r > 0$ we have:

$$\left(1 - \frac{2\Phi^{-1}(\frac{1}{2}+\frac{q}{2})\phi(\Phi^{-1}(\frac{1}{2}+\frac{q}{2}))}{q}\right) \leq \frac{q}{8}\frac{\Phi^{-1}\left(\frac{1}{2}+\frac{q}{2}\right)}{\phi\left(\Phi^{-1}\left(\frac{1}{2}+\frac{q}{2}\right)\right)} \leq \frac{q}{8}\frac{r}{\phi(r)}.$$

□

Proof of Proposition 3.2.13. First note that if Y has a bounded support, this propositionosition clearly follows from assumption (3.2.9) alone. Thus, without loss of generality we assume that Y has unbounded support (from both sides, as if one of them is bounded we can handle it in much the same way as the proof below).

Let $\widetilde{B}_0 = B_0 + \eta$, for some small fixed $\eta > 0$. Fix any partition $a \in \mathcal{A}_H(l, K)$. Let $S_0 = \{h : a_h \in [-\widetilde{B}_0, \widetilde{B}_0]\}$, and let $h_m = \min S_0, h_M = \max S_0$. Note that the following simple inequality holds for any $h \geq 2, h \leq h_m - 2$ or $h \geq h_M + 1, h \leq H - 1$:

$$\begin{aligned}\mathrm{Var}[m(Y)|a_h < Y \leq a_{h+1}] &\leq \inf_{t\in(a_h,a_{h+1}]} \mathbb{E}[(m(Y) - m(t))^2|a_h < Y \leq a_{h+1}] \\ &\leq \sup_{y,t\in(a_h,a_{h+1}]} (m(y) - m(t))^2 \\ &\leq (\widetilde{m}(|a_h|) - \widetilde{m}(|a_{h+1}|))^2.\end{aligned}$$

This gives us the following inequality:

$$\begin{aligned}\sum_{h=2}^{h_m-2} \mathrm{Var}[m(Y)|a_h < Y \leq a_{h+1}] &\leq \sum_{h=2}^{h_m-2} (\widetilde{m}(|a_h|) - \widetilde{m}(|a_{h+1}|))^2 \\ &\leq (\widetilde{m}(|a_2|) - \widetilde{m}(|a_{h_m-1}|))^2,\end{aligned} \tag{B.1.1}$$

where the last inequality holds since $\widetilde{m}$ is non-decreasing. Similar inequality holds for the other tail as well.

Using a similar technique we get the following bound on the interval: $[-\widetilde{B}_0, \widetilde{B}_0]$:

$$\sum_{h=h_m}^{h_M-1} \mathrm{Var}[m(Y)|a_h < Y \le a_{h+1}] \le \sum_{h=h_m}^{h_M-1} \mathbb{E}[(m(Y) - m(a_h))^2 | a_h < Y \le a_{h+1}]$$
$$\le \sum_{h=h_m}^{h_M-1} \sup_{y \in (a_h, a_{h+1}]} (m(y) - m(a_h))^2.$$

Notice further that:

$$\mathrm{Var}[m(Y)|a_{h_m-1} < Y \le a_{h_m}] \le \sup_{y \in (a_{h_m-1}, a_{h_m}]} (m(y) - m(-\widetilde{B}_0))^2$$
$$\le \sup_{y \in (a_{h_m-1}, -\widetilde{B}_0]} (m(y) - m(-\widetilde{B}_0))^2 + \sup_{y \in [-\widetilde{B}_0, a_{h_m}]} (m(y) - m(-\widetilde{B}_0))^2.$$

And a similar inequality holds for $\mathrm{Var}[m(Y)|a_{h_M} < Y \le a_{h_M+1}]$. Thus:

$$\sum_{h=h_m-1}^{h_M} \mathrm{Var}[m(Y)|a_h < Y \le a_{h+1}] \le \underbrace{\sum_{h=h_m}^{h_M-1} \sup_{y \in (a_h, a_{h+1}]} (m(y) - m(a_h))^2}_{I_1}$$
$$+ \underbrace{\sup_{y \in (a_{h_m-1}, -\widetilde{B}_0]} (m(y) - m(-\widetilde{B}_0))^2}_{I_2} + \underbrace{\sup_{y \in [-\widetilde{B}_0, a_{h_m}]} (m(y) - m(-\widetilde{B}_0))^2}_{I_3}$$
$$+ \underbrace{\sup_{y \in [\widetilde{B}_0, a_{h_M+1}]} (m(y) - m(\widetilde{B}_0))^2}_{I_4} + \underbrace{\sup_{y \in (a_{h_M}, \widetilde{B}_0]} (m(y) - m(\widetilde{B}_0))^2}_{I_5}.$$

We have:

$$I_1 + I_3 + I_5 \leq \sup_{b \in \Pi_{2|S_0|+3}(\widetilde{B}_0)} \sum_{i=2}^{2|S_0|+3} (m(b_i) - m(b_{i-1}))^2 \tag{B.1.2}$$
$$\leq \sup_{b \in \Pi_{2|S_0|+3}(\widetilde{B}_0)} \left(\sum_{i=2}^{2|S_0|+3} |m(b_i) - m(b_{i-1})| \right)^2 .$$

To see this, consider a partition containing the points $b_1 = -\widetilde{B}_0, b_3 = a_{h_m}, \ldots, b_{2|S_0|+1} = a_{h_M}, b_{2|S_0|+3} = \widetilde{B}_0$, and $b_{2k} = \operatorname{argmax}_{y \in (b_{2k-1}, b_{2k+1}]} (m(y) - m(b_{2k-1}))^2$ (note that if the max doesn't exist we can take a limit of partitions converging to it).

Next, we control I_2:

$$I_2 = \sup_{y \in (a_{h_m-1}, -\widetilde{B}_0]} (m(y) - m(-\widetilde{B}_0))^2 \leq (\widetilde{m}(a_{h_m-1}) - \widetilde{m}(\widetilde{B}_0))^2.$$

with the last inequality following from (3.2.10). Combining this bound with (B.1.1) we get:

$$(\widetilde{m}(|a_2|) - \widetilde{m}(|a_{h_m-1}|))^2 + I_2 \leq (\widetilde{m}(|a_2|) - \widetilde{m}(\widetilde{B}_0))^2. \tag{B.1.3}$$

Similarly, for I_4 and the other bound in (B.1.1) we have:

$$(\widetilde{m}(|a_H|) - \widetilde{m}(|a_{h_M+1}|))^2 + I_4 \leq (\widetilde{m}(|a_H|) - \widetilde{m}(\widetilde{B}_0))^2. \tag{B.1.4}$$

Finally, we deal with the tail part:

$$\begin{aligned}
\mathrm{Var}[m(Y)|Y \le a_2] &\le \mathbb{E}[(m(Y) - m(a_2))^2 | Y \le a_2] \\
&\le \mathbb{E}[(\widetilde{m}(|Y|) - \widetilde{m}(|a_2|))^2 | Y \le a_2] \\
&\le 4\mathbb{E}[(\widetilde{m}(|Y|))^2 | Y \le a_2] \\
&\le 4(\mathbb{E}[|\widetilde{m}(|Y|)|^{2+\xi} | Y \le a_2])^{2/(2+\xi)} \\
&= 4\left(\int_{-\infty}^{a_2} |\widetilde{m}(|y|)|^{2+\xi} d\mathbb{P}(Y \le y)\mathbb{P}(Y \le a_2)^{-1}\right)^{2/(2+\xi)} \\
&= o(1)\mathbb{P}(Y \le a_2)^{-2/(2+\xi)}.
\end{aligned} \tag{B.1.5}$$

where we used the fact that $\mathbb{E}[|\widetilde{m}(|Y|)|^{2+\xi}]$ is bounded by assumption, and the $o(1)$ is in the sense of $|a_2| \to \infty$. We can show a similar inequality for the other tail — $\mathrm{Var}[m(Y)|Y \ge a_H]$.

Combining (B.1.1), (B.1.3), (B.1.4), (B.1.2) and (B.1.5) we have:

$$\begin{aligned}
&\sum_{h=1}^{H} \mathrm{Var}[m(Y)|a_h < Y \le a_{h+1}] \le \sup_{b \in \Pi_{2|S_0|+3}(\widetilde{B}_0)} \left(\sum_{i=2}^{2|S_0|+3} |m(b_i) - m(b_{i-1})|\right)^2 \\
&+ o(1)\mathbb{P}(Y \ge a_H)^{-2/(2+\xi)} + o(1)\mathbb{P}(Y \le a_2)^{-2/(2+\xi)} \\
&+ (\widetilde{m}(|a_2|) - \widetilde{m}(\widetilde{B}_0))^2 + (\widetilde{m}(|a_H|) - \widetilde{m}(\widetilde{B}_0))^2.
\end{aligned}$$

Since $(\widetilde{m}(|a_2|) - \widetilde{m}(\widetilde{B}_0))^2 \le 4(\widetilde{m}(|a_2|))^2$, and we know that $\widetilde{m}(|a_2|)^{2+\xi} \frac{l}{H} \le \widetilde{m}(|a_2|)^{2+\xi}\mathbb{P}(Y \le a_2) \to 0$, this means that $\widetilde{m}(|a_2|)^2 \frac{1}{H^{2/(2+\xi)}} \to 0$. Furthermore $o(1)\mathbb{P}(Y \le a_2)^{-2/(2+\xi)} \frac{1}{H^{2/(2+\xi)}} = o(1)$. Finally we recall that by (3.2.9) we have:

$$\sup_{b\in\Pi_{2|S_0|+3}(\widetilde{B}_0)}\left(\sum_{i=2}^{2|S_0|+3}|m(b_i)-m(b_{i-1})|\right)^2 \leq \left(\sup_{b\in\Pi_{2|S_0|+3}(\widetilde{B}_0)}\sum_{i=2}^{2|S_0|+3}|m(b_i)-m(b_{i-1})|\right)^2$$
$$= o\left(|S_0|^{2/(2+\xi)}\right).$$

However $|S_0| \leq \mathbb{P}(-\widetilde{B}_0 < Y < \widetilde{B}_0)H/l + 1$ and thus:

$$\sup_{b\in\Pi_{2|S_0|+3}(\widetilde{B}_0)}\left(\sum_{i=2}^{2|S_0|+3}|m(b_i)-m(b_{i-1})|\right)^2 = o\left(H^{2/(2+\xi)}\right),$$

which finishes the proof. □

Proof of Lemma 3.4.1. Before we go to the main proof of the lemma we first formulate a key result, which enables us to prove this lemma. We note that this result might be of independent interest.

Lemma B.1.1. *Let $A(X,\nu) \in \{0,1\}$ be any acceptance rule such that $\mathbb{P}(A=1) \geq q$, where $X \sim N(0,1)$ and ν be any random variable. Let $X_1,\ldots,X_n$ be an iid samples of the distribution $X|A(X,\nu)=1$. Denote with $\mu = \mathbb{E}[X_i]$. Then we have:*

$$\mathbb{P}(|\overline{X}-\mu| > \epsilon)$$
$$\leq \inf_{M>0} 2\exp\left(-\frac{1}{2}\frac{\epsilon^2 n}{M\epsilon + 2\exp\left(\frac{2}{M^2}\right)\left[M^2 + 2\kappa M\sqrt{-\log\left(\frac{q}{2}\right)} + (2\kappa)^2 C_{\frac{2\kappa}{M}}\left(-\log\left(\frac{q}{2}\right)\frac{q}{2}\right)\right]}\right),$$

where $\kappa = \frac{\phi(\Phi^{-1}(1-\frac{q}{2}))}{\sqrt{-\log(\frac{q}{2})\frac{q}{2}}}$, and $C_r = \frac{\exp(r)-1-r}{r^2}$.

Remark B.1.2. *The constant κ here can be shown to be $\leq \sqrt{2}r$ for all values q satisfying $\frac{q}{2} \leq 1 - \Phi\left(\frac{1}{\sqrt{r-1}}\right)$.*

We assumed that H and ϵ are specified so that (quite arbitrary) $\frac{\frac{1}{H}-2\epsilon}{2} < 1 - \Phi(1/\sqrt{\sqrt{2}-1})$, so we can select $\kappa = 2$ and we select $M = 1$. By (3.4.3) we know that $\mathbb{P}(Y \in S_h) \geq \frac{1}{H} - 2\epsilon$ on S, thus setting $q = \frac{1}{H} - 2\epsilon$, by Lemma B.1.1 conditionally on $\{Y_{(mh)} : h = 1, \ldots, H-1\}$ we have for all $j \in S_\beta$ and all h:

$$\begin{aligned}
&\mathbb{P}\left(\left|\overline{X}^j_{h,1:(m-1)} - \mu^j_h\right| > \eta\right) \\
&\leq 2\exp\left(-\frac{1}{2}\frac{\eta^2(m-1)}{\eta + 2\exp(2)\left[1 + 4\sqrt{-\log\left(\frac{q}{2}\right)} + 16C_4\left(-\log\left(\frac{q}{2}\right)\frac{q}{2}\right)\right]}\right) \\
&= 2\exp\left(-\frac{1}{2}\frac{\eta^2(m-1)}{\eta + \widetilde{C}_1 + \widetilde{C}_2\sqrt{-\log\left(\frac{q}{2}\right)} + \widetilde{C}_3\left(-\log\left(\frac{q}{2}\right)q\right)}\right),
\end{aligned}$$

where $\widetilde{C}_1 = 2\exp(2)$, $\widetilde{C}_2 = 8\exp(2)$ and $\widetilde{C}_3 = 32C_4\exp(2)$ are absolute constants.

Note that Lemma B.1.1 is applicable in this case, as the statistics $X^j_{h,i}$ are conditionally independent on $Y_{(m(h-1))}$ and $Y_{(mh)}$ as we noticed when we described the second data generation procedure in the main text, and therefore we can set the acceptance rule $A(X, \varepsilon) := \mathbb{1}(f(\beta^\intercal X, \varepsilon) \in S_h)$. Furthermore, notice that the above inequality holds regardless of the values of $\{Y_{(mh)} : h = 1, \ldots, H-1\}$, on the event S.

Finally, using union bound across the slices and the indexes $j \in S_\beta$, we have that this holds for all slices or in other words

$$\begin{aligned}
&\mathbb{P}\left(\max_{j\in S_\beta, h\in\{1,\ldots,H\}}\left|\overline{X}^j_{h,1:(m-1)} - \mu^j_h\right| > \eta\right) \leq \\
&2sH\exp\left(-\frac{1}{2}\frac{\eta^2(m-1)}{\eta + \widetilde{C}_1 + \widetilde{C}_2\sqrt{-\log\left(\frac{q}{2}\right)} + \widetilde{C}_3\left(-\log\left(\frac{q}{2}\right)q\right)}\right),
\end{aligned}$$

on the event S. This is precisely what we wanted to show. □

Proof of Lemma B.1.1. Before we go to the main proof of the lemma let's consider the following simple proposition, which is the key to show the bound:

Proposition B.1.3. *Let $A(X,\nu)$ be any acceptance rule with $\mathbb{P}(A(X,\nu)) \geq q$, and M be any fixed constant, and $X \sim N(0,1)$. Consider the random variable $\widetilde{X} = [X|A(X,\nu)=1]$. Then we have:*

$$\mathbb{E}[\exp(|\widetilde{X}|/M)] \leq \exp\left(\frac{1}{2M^2}\right)\left[\frac{\frac{q}{2} + \frac{1}{M}\phi(\Phi^{-1}(1-\frac{q}{2})) - \frac{1}{2M^2}\phi^2(\Phi^{-1}(1-\frac{q}{2})) + \ldots)}{\frac{q}{2}}\right].$$

Proof of Proposition B.1.3. We first show that $\mathbb{E}[\exp(|\widetilde{X}|/M)] \leq \mathbb{E}[\exp(|X|/M)||X| > \Phi^{-1}(1 - q/2)]$. Clearly we have:

$$\mathbb{E}[\exp(|\widetilde{X}|/M)] = \sum_{i=0}^{\infty} \frac{\mathbb{E}[(|\widetilde{X}|/M)^i]}{i!}.$$

Note that, $\mathbb{P}(|\widetilde{X}| \geq t) \leq \frac{\mathbb{P}(|X|\geq t, A(X,\nu)=1)}{q} \leq \frac{2-2\Phi(t)}{q}$. Note that the last estimate is trivial when $t \leq \Phi^{-1}(1-\frac{q}{2})$ (i.e. the RHS is bigger than 1).

Now using the following well known formula, for $i \geq 1$:

$$\begin{aligned}\mathbb{E}[|\widetilde{X}|^i] = \int_0^{\infty} \mathbb{P}(|\widetilde{X}| \geq t) i t^{i-1} dt &\leq \int_0^{\Phi^{-1}(1-\frac{q}{2})} i t^{i-1} dt + \int_{\Phi^{-1}(1-\frac{q}{2})}^{\infty} \frac{2-2\Phi(t)}{q} i t^{i-1} dt \\ &= \mathbb{E}[|X|^i||X| > \Phi^{-1}(1-q/2)].\end{aligned}$$

Where we applied the expectation formula in the last expression again. Finally summing up over i, gives the desired result:

$$\begin{aligned}\sum_{i=0}^{\infty} \frac{\mathbb{E}[(|\widetilde{X}|/M)^i]}{i!} &\leq \sum_{i=0}^{\infty} \frac{\mathbb{E}[(|X|/M)^i||X| > \Phi^{-1}(1-q/2)]}{i!} \\ &= \mathbb{E}[\exp(|X|/M)||X| > \Phi^{-1}(1-q/2)].\end{aligned}$$

Note that we have, $\mathbb{E}[\exp(|X|/M)||X| > \Phi^{-1}(1-q/2)] = \mathbb{E}[\exp(X/M)|X > \Phi^{-1}(1-q/2)]$. Using the mgf of a truncated normal distribution, we get the following:

$$\mathbb{E}[\exp(X/M)|X > \Phi^{-1}(1-q/2)] = \exp\left(\frac{1}{2M^2}\right)\left[\frac{1-\Phi(\Phi^{-1}(1-\frac{q}{2})-\frac{1}{M})}{\frac{q}{2}}\right].$$

The propositionosition now follows after a Taylor expansion. □

By the definition of κ we have $\phi(\Phi^{-1}(1-\frac{q}{2})) = \kappa\sqrt{-\log(\frac{q}{2})}\frac{q}{2}$. Therefore we have:

$$\begin{aligned}
&\mathbb{E}[\exp(|\widetilde{X}|/M)] \\
&\leq \exp\left(\frac{1}{2M^2}\right)\left[\frac{\frac{q}{2}+\frac{1}{M}\phi(\Phi^{-1}(1-\frac{q}{2}))+\sum_{i=2}^{\infty}\frac{\kappa^i\left(\sqrt{-\log(\frac{q}{2})}\frac{q}{2}\right)^i)}{i!M^i}}{\frac{q}{2}}\right] \\
&= \exp\left(\frac{1}{2M^2}\right)\left[\frac{\frac{q}{2}+\frac{1}{M}\phi(\Phi^{-1}(1-\frac{q}{2}))+\exp\left(\frac{\kappa}{M}\sqrt{-\log(\frac{q}{2})}\frac{q}{2}\right)-1-\frac{\kappa}{M}\sqrt{-\log(\frac{q}{2})}\frac{q}{2}}{\frac{q}{2}}\right] \\
&\leq \exp\left(\frac{1}{2M^2}\right)\left[\frac{\frac{q}{2}+\frac{1}{M}\phi(\Phi^{-1}(1-\frac{q}{2}))+C_{\frac{\kappa}{M}}\left(\frac{\kappa}{M}\sqrt{-\log(\frac{q}{2})}\frac{q}{2}\right)^2}{\frac{q}{2}}\right],
\end{aligned}$$

where we used the fact that $\exp(x)-1-x \leq C_r x^2$ for all $0 < x \leq r$ where $C_r = \frac{\exp(r)-1-r}{r^2}$, which can be checked easily by noting that C_r is an increasing function of r, and further that $\sqrt{-\log(x)}x < 1$ for all $0 < x \leq 1$. So what this discussion gives us is that:

$$\mathbb{E}[\exp(|\widetilde{X}|/M)] \leq \exp\left(\frac{1}{2M^2}\right)\left[1+\frac{\kappa}{M}\sqrt{-\log\left(\frac{q}{2}\right)}+\left(\frac{\kappa}{M}\right)^2 C_{\frac{\kappa}{M}}\left(-\log\left(\frac{q}{2}\right)\frac{q}{2}\right)\right].$$

Recall that a version of Bernstein's inequality requires the following moment condition: $\mathbb{E}[|Z|^m] \leq m!\frac{M^{m-2}v}{2}$ for $m \geq 2$, (e.g. see[82] Lemma 2.2.11). By a Taylor expansion it can be easily seen that this condition is implied by $\mathbb{E}[\exp(|Z|/M)-1-|Z|/M]M^2 \leq v/2$.

Obviously we have $\mathbb{E}[\exp(|Z|/M) - 1 - |Z|/M]M^2 \leq \mathbb{E}[\exp(|Z|/M)]M^2$ and therefore if we can find a v such that $v \geq 2\mathbb{E}[\exp(|Z|/M)]M^2$ we will be able to apply Bernstein's inequality. Note that for our random variables we have:

$$\begin{aligned}\mathbb{E}[\exp(|\widetilde{X} - \mu|/M)]M^2 &\leq \mathbb{E}[\exp(|\widetilde{X}|/M + |\mu|/M)]M^2 \\ &\leq \mathbb{E}[\exp(2|\widetilde{X}|/M)]M^2.\end{aligned}$$

The last inequality following from a double Jensen's inequality, upon noticing that the function $\exp(|\cdot|)$ is convex, and then putting the square inside the expectation.

$$\mathbb{E}[\exp(2|\widetilde{X}|/M)]M^2 \leq \exp\left(\frac{2}{M^2}\right)\left[M^2 + 2\kappa M\sqrt{-\log\left(\frac{q}{2}\right)} + (2\kappa)^2 C_{\frac{2\kappa}{M}}\left(-\log\left(\frac{q}{2}\right)\frac{q}{2}\right)\right].$$

Finally applying Bersntein's inequality and taking $\inf$ with respect to M gives the desired result. □

Proof of Remark B.1.2. We show the remark here. Using the well known fact that for all $x > 0$: $\phi(x) \leq (1 - \Phi(x))\left(x + \frac{1}{x}\right)$, we have that $\phi(x) \leq r(1 - \Phi(x))x$, for $x \geq \frac{1}{\sqrt{r-1}}$. Thus for values of $q/2 \leq 1 - \Phi(\sqrt{\frac{1}{r-1}}), r > 1$, we have:

$$\frac{\phi(\Phi^{-1}(1 - \frac{q}{2}))}{\sqrt{-\log(\frac{q}{2})\frac{q}{2}}} \leq r\frac{\Phi^{-1}\left(1 - \frac{q}{2}\right)}{\sqrt{-\log(q/2)}} \leq \sqrt{2}r,$$

where the last inequality follows from $q/2 = 1 - \Phi(x) \leq \exp(-x^2/2)$ (for $x \geq 0$, or equivalently $q/2 \leq \frac{1}{2}$, but we obviously have $q/2 \leq 1 - \Phi\left(\frac{1}{\sqrt{r-1}}\right) < \frac{1}{2}$), and thus $\Phi^{-1}(1 - q/2) = x \leq \sqrt{2}\sqrt{-\log(q/2)}$. □

Proof of Lemma 3.4.2. Using the sliced stable condition, for large H we get:

$$\begin{aligned}
&\left|\mathrm{Var}[m_j(Y)] - \sum_{i=1}^{n}(\mu_h^j)^2\mathbb{P}(Y \in S_h)\right| \\
&= \sum_{i=1}^{n}\mathrm{Var}[m_j(Y)|Y \in S_h]\mathbb{P}(Y \in S_h) \\
&\leq \frac{CH^\kappa}{s}\left(\frac{1}{H} + 2\epsilon\right).
\end{aligned}$$

This shows (3.4.5). Consequently we have:

$$\begin{aligned}
\left(\frac{1}{H} - 2\epsilon\right)\sum_{h=1}^{H}(\mu_h^j)^2 &\leq \sum_{h=1}^{H}(\mu_h^j)^2\mathbb{P}(Y \in S_h) \\
&\leq \frac{C_V}{s} + \frac{CH^\kappa}{s}\left(\frac{1}{H} + 2\epsilon\right).
\end{aligned}$$

This yields (3.4.6). To get (3.4.7) we proceed as follows:

$$\begin{aligned}
&\left(\frac{1}{H} - 2\epsilon\right)\sum_{h=1}^{H}|\mu_h^j| \leq \sum_{h=1}^{H}|\mu_h^j|\mathbb{P}(Y \in S_h) \\
&\leq \sqrt{\sum_{h=1}^{H}\mathbb{P}(Y \in S_h)}\sqrt{\sum_{h=1}^{H}(\mu_h^j)^2\mathbb{P}(Y \in S_h)} \\
&\leq \sqrt{\frac{C_V}{s} + \frac{CH^\kappa}{s}\left(\frac{1}{H} + 2\epsilon\right)},
\end{aligned}$$

and we are done. □

Proof of Lemma 3.4.3. Note that on the event $\widetilde{S}$ we have the following chain of inequalities:

$$\begin{aligned}
&\frac{1}{H}\sum_{h=1}^{H}\left|\left(\frac{1}{m}X^j_{h,m}+\frac{m-1}{m}\overline{X}^j_{h,1:(m-1)}\right)^2-\frac{(m-1)^2}{m^2}(\mu^j_h)^2\right| \qquad \text{(B.1.6)}\\
&\leq \frac{1}{Hm^2}\sum_{h=1}^{H}(X^j_{h,m})^2+\frac{2(m-1)}{Hm^2}\sum_{h=1}^{H}|X^j_{h,m}|(\eta+|\mu^j_h|)\\
&+\frac{1}{H}\frac{(m-1)^2}{m^2}\sum_{h=1}^{H}\eta(2|\mu^j_h|+\eta)\\
&\leq \frac{1}{m}\frac{1}{n}\sum_{r=1}^{n}(X^j_r)^2+\frac{2}{Hm}\sqrt{\sum_{r=1}^{n}(X^j_r)^2}\sqrt{2\sum_{h=1}^{H}(\eta^2+(\mu^j_h)^2)}\\
&+\eta^2+2\frac{\eta}{H}B_3.
\end{aligned}$$

where we used that we are on the event $\widetilde{S}$ in the first inequality, and (3.4.7), Cauchy-Schwartz and the trivial bounds $\frac{m-1}{m}<1$, $\frac{1}{n}\sum_{h=1}^{H}(X^j_{h,m})^2\leq\frac{1}{n}\sum_{r=1}^{n}(X^j_r)^2\sim\chi^2_n/n$ in the second one.

Using a χ^2 tail bound provided in[37], we have:

$$\mathbb{P}(\frac{1}{n}\chi^2_n>1+\tau)<\exp(-\frac{3}{16}n\tau^2),\ \tau\in[0,\frac{1}{2}).$$

Hence we infer that there exists a set $\widetilde{\widetilde{S}}\subset\widetilde{S}$ failing with probability at most $s\exp(-\frac{3}{16}n\tau^2)$, such that $\frac{1}{n}\sum_{r=1}^{n}(X^j_r)^2\leq 1+\tau$ for all $j\in S_\beta$. Therefore continuing the bound on the event $\widetilde{\widetilde{S}}$, we get:

$$(B.1.6)\leq\frac{(1+\tau)}{m}+\frac{2\sqrt{1+\tau}}{\sqrt{m}}\sqrt{2\eta^2+2\frac{B_2}{H}}+\eta^2+2\eta\frac{B_3}{H},$$

where we used (3.4.6). This finishes the proof. □

Proof of Corollary 3.2.4. Note that we can clearly rewrite $\widehat{V}^{jj}=\frac{1}{H}\sum_{h=1}^{H}(\overline{X}^j_h-\mu^j)^2-(\overline{X}^j-\mu^j)^2$.

For the first term we can use the proof Theorem 3.2.3 to conclude that $V^{jj} \geq \frac{C_V}{2s}$ for $j \in S_\beta$ and $V^{jj} \leq \frac{C_V}{4s}$ for $j \in S_\beta^c$ on an event with asymptotic probability 1. Next we show that $(\overline{X}^j - \mu^j)^2$ is asymptotically negligible. Clearly $\overline{X}^j - \mu^j \sim N(0, n^{-1})$. Hence by a standard normal tail bound $\mathbb{P}(|\overline{X}^j - \mu^j| \geq x) \leq 2\exp(-nx^2/2)$, and thus by a union bound:

$$\mathbb{P}\left(\max_{j \in S_\beta} |\overline{X}^j - \mu^j| \geq \sqrt{2\frac{\log(p-s)}{n}}\right) \leq 2s(p-s)^{-1} = o(1).$$

Thus $\max_{j \notin S_\beta}(\overline{X}^j - \mu^j)^2 \leq 2\frac{\log(p-s)}{n} \leq \frac{2}{\Omega s}$ with probability not smaller than $1 - 2s(p-s)^{-1}$. Hence if $\Omega > 16C_V^{-1}$ e.g., we will have $\widehat{V}^{jj} \geq \frac{3C_V}{8s}$ for $j \in S_\beta$, while $\widehat{V}^{jj} \leq V^{jj} \leq \frac{C_V}{4s}$ for $j \notin S_\beta$ with probability converging to 1, which completes the proof. □

Proof of Lemma 3.5.2. We first note that the following inequality holds:

$$\begin{aligned}
&\left|V^{jk} - \operatorname{sign}(\beta_j)\operatorname{sign}(\beta_k)\frac{C_V}{s}\right| = \left|\frac{1}{H}\sum_{h=1}^{H}\overline{X}_h^j\overline{X}_h^k - \operatorname{sign}(\beta_j)\operatorname{sign}(\beta_k)\frac{C_V}{s}\right| \\
&\leq \left|\frac{1}{H}\sum_{h=1}^{H}(\overline{X}_h^j)^2 - \frac{C_V}{s}\right| + \frac{1}{H}\sum_{h=1}^{H}\left|\overline{X}_h^j\right|\left|\operatorname{sign}(\beta_j)\overline{X}_h^j - \operatorname{sign}(\beta_k)\overline{X}_h^k\right|,
\end{aligned}$$

where the LHS equals, the LHS of (3.5.3) after using (3.5.2). Fortunately (3.4.8) and (3.4.9) already give bounds on the first term on the event $\widetilde{\widetilde{S}}$. We now show that the second term is small on the same event. Note that the following identity holds:

$$\begin{aligned}
&\frac{1}{H}\sum_{h=1}^{H}\left|\overline{X}_h^j\right|\left|\operatorname{sign}(\beta_j)\overline{X}_h^j - \operatorname{sign}(\beta_k)\overline{X}_h^k\right| \\
&= \frac{1}{H}\sum_{h=1}^{H}\left|\frac{m-1}{m}\overline{X}_{h,1:(m-1)}^j + \frac{1}{m}X_{h,m}^j\right|\left|\operatorname{sign}(\beta_j)\overline{X}_h^j - \frac{m-1}{m}\operatorname{sign}(\beta_k)\mu_h^j\right. \\
&\qquad\qquad \left. + \frac{m-1}{m}\operatorname{sign}(\beta_k)\mu_h^k - \operatorname{sign}(\beta_k)\overline{X}_h^k\right|.
\end{aligned}$$

Thus on the event $\widetilde{\widetilde{S}}$:

$$\frac{1}{H}\sum_{h=1}^{H}\left|\overline{X}_h^j\right|\left|\operatorname{sign}(\beta_j)\overline{X}_h^j-\operatorname{sign}(\beta_k)\overline{X}_h^k\right|$$
$$\leq \frac{1}{H}\sum_{h=1}^{H}\left(\mu_h+\eta+\frac{1}{m}\left|X_{h,m}^j\right|\right)\left(2\eta+\frac{1}{m}\left|X_{h,m}^j\right|+\frac{1}{m}\left|X_{h,m}^k\right|\right),$$

where $\mu_h=|\mu_h^j|=|\mu_h^k|$, and we used that $\frac{m-1}{m}<1$, and the fact that on $\widetilde{\widetilde{S}}$ we have $|X_{h,1:(m-1)}^j-\mu_h^j|\leq\eta$ and similarly $|X_{h,1:(m-1)}^k-\mu_h^k|\leq\eta$. Next we have:

$$\frac{1}{H}\sum_{h=1}^{H}\left(\mu_h+\eta+\frac{1}{m}\left|X_{h,m}^j\right|\right)\left(2\eta+\frac{1}{m}\left|X_{h,m}^j\right|+\frac{1}{m}\left|X_{h,m}^k\right|\right)$$
$$\leq 2\frac{\eta}{H}\sum_{h=1}^{H}\mu_h+\frac{1}{Hm}\sum_{h=1}^{H}(\mu_h+\eta)(|X_{h,m}^j|+|X_{h,m}^k|)+2\eta^2$$
$$+\frac{2\eta}{mH}\sum_{h=1}^{H}|X_{h,m}^j|+\frac{1}{m^2H}\sum_{h=1}^{H}(X_{h,m}^j)^2+\frac{1}{2m^2H}\sum_{h=1}^{H}(X_{h,m}^j)^2+\frac{1}{2m^2H}\sum_{h=1}^{H}(X_{h,m}^k)^2,$$

where we used the simple inequality $ab\leq(a^2+b^2)/2$. Luckily we have already controlled all of the above quantities. Using Lemma 3.4.3 and (3.4.8) we get:

$$\frac{1}{H}\sum_{h=1}^{H}\left|\overline{X}_h^j\right|\left|\operatorname{sign}(\beta_j)\overline{X}_h^j-\operatorname{sign}(\beta_k)\overline{X}_h^k\right|$$
$$\leq 2\frac{\eta}{H}B_3+\frac{2\sqrt{1+\tau}}{\sqrt{m}}\sqrt{2\eta^2+2\frac{B_2}{H}}+2\eta^2$$
$$+2\frac{\eta\sqrt{1+\tau}}{\sqrt{m}}+2\frac{1+\tau}{m},$$

where we heavily relied on the fact that on $\widetilde{\widetilde{S}}$ we have $\frac{1}{mH}\sum_{r=1}^{n}(X_r^j)^2\leq 1+\tau$, the rest of the bounds can be seen in the proof of Lemma 3.4.3. (For the term note $\frac{1}{mH}\sum_{h=1}^{H}|X_{h,m}^j|\leq\frac{1}{m}\sqrt{\sum_{h=1}^{H}(X_{h,m}^j)^2/H}\leq\frac{1}{\sqrt{m}}\sqrt{\sum_{r=1}^{n}(X_r^j)^2/mH}$).

Finally noting that $2\frac{\eta\sqrt{1+\tau}}{\sqrt{m}} \leq \eta^2 + \frac{1+\tau}{m}$ gives the desired result. □

Proof of Corollary 3.2.6. Note that we can rewrite $\widehat{V}^{jk} = \frac{1}{H}\sum_{h=1}^{H}(\overline{X}_h^j - \mu^j)(\overline{X}_h^k - \mu^k) - (\overline{X}^j - \mu^j)(\overline{X}^k - \mu^k)$. Similarly to the proof of Corollary 3.2.4 we have $\mathbb{P}(\max_{j\in S_\beta}|\overline{X}^j - \mu^j| \geq x) \leq 2s\exp(-nx^2/2)$. Then we have $\max_{j\in S_\beta}|\overline{X}^j - \mu^j| \leq \sqrt{\frac{\log(st)}{n}}$ with probably no less than $1 - 2t^{-1}$. This implies that:

$$\max_{j\in S_\beta}(\overline{X}^j - \mu^j)^2 \leq \frac{\log(st)}{n} \leq \Omega^{-1}\frac{\log s + \log t}{s\log(p-s)}.$$

Therefore using bound (3.5.8), we get:

$$\sup_{j,k\in S_\beta}\left|\widehat{V}^{jk} - \mathrm{Cov}(m_j(Y), m_k(Y))\right| \leq \underbrace{\frac{C_V}{2st} + \frac{\Omega^{-1}\log(st)}{s\log(p-s)}}_{B(s)}.$$

It is easily seen that the fact that $\log s = o(\log p)$ implies that $sB(s) = o(1)$, and hence all bounds from the first part of the proof of Theorem 3.2.5 hold in this setting as well.

Next we move on to show that the matrix U, defined in (3.5.1), can be selected so that the blocks $\widetilde{V}_{S_\beta^c S_\beta}, \widetilde{V}_{S_\beta^c S_\beta^c}$ are 0, where $\widetilde{V} = \widehat{V} - \frac{C_V}{2s}U$. Note that since $\widehat{V}^{kk} \leq V^{kk}$ for all $k \in S_\beta^c$, we will just show the bound for $\widehat{V}^{jj}, j \in S_\beta$. We have:

$$\widehat{V}^{jj} \leq \frac{C_V}{s} + B(s) \leq \frac{3C_V}{2s},$$

with the last inequality holding asymptotically as we saw before. Hence the rest of the proof of Theorem 3.2.5 is valid and we are done. □

Proof of Theorem 3.2.7. In this proof we show the lower bound on n, such that detecting the support of β is impossible, with probability at least $\frac{1}{2}$. We follow closely the approaches showed in[3], and rely on using Fano's inequality, which in turn is a standard approach for showing minimax

lower bounds (e.g. see[16,86,90,92] among others). For simplicity of the exposition we will assume that the vector β has only non-negative entries (i.e. all non-zero entries are $\frac{1}{\sqrt{s}}$). The proof extends in exactly the same way in the case when the entries of β are not restricted to be positive.

As we saw in Section 3.2.1, the space of models satisfying conditions (3.2.1) and (3.2.2), include models of the form $Y = f(\beta^\intercal X + \varepsilon)$, where f is a monotone function, and $\varepsilon \sim N(0, \sigma^2)$. Note that if σ^2 is specified in such a manner that $C_V = \frac{1}{1+\sigma^2}$, we have that $f = Id$, satisfies condition (3.2.2). Hence, our conditions include the simple linear regression with Gaussian noise as a subset. The lower bound then cannot be bigger than the one for the linear regression model. A lower bound on support recovery for sparse linear models can be found in[86] in a more generic setup than what we consider here, but we present a simpler proof with a slightly better constant, for completeness.

Let $[p] = \{1, \dots, p\}$. Denote with $\mathbb{S} \subset 2^{[p]}$, the set of all subsets of $[p]$ with s elements. Clearly, $|\mathbb{S}| = \binom{p}{s}$. Let $\widehat{S} : (\mathbb{R}^{p+1})^n \to \mathbb{S}$ be any potentially random function, which is used to recover the support of β, based on the sample $\{(Y_i, X_i)\}_{i=1}^n$. Under the 0-1 loss the risk equals the probability of error:

$$\frac{1}{\binom{p}{s}} \sum_{S_\beta \in \mathbb{S}} P_{S_\beta}(\widehat{S} \neq S_\beta), \tag{B.1.7}$$

where by P_{S_β} we are measuring the probability under a dataset generated with $\text{supp}(\beta)$ equal to the index of the measure P_{S_β}.

Instead of directly dealing with the sum above, we first consider the $p - s + 1$ element set $\widetilde{\mathbb{S}} = \{S \in \mathbb{S} : \{1, \dots, s-1\} \subset S\}$, and we bound the probability of error, on any function $\widehat{S}$ (even if given the knowledge that the true support is drawn from $\widetilde{\mathbb{S}}$). If U is a uniformly selected subset of $\widetilde{\mathbb{S}}$

we then have by Fano's inequality that:

$$\mathbb{P}(\text{error}) \geq 1 - \frac{I(U;(Y,X)^n) + \log(2)}{\log|\widetilde{\mathbb{S}}|},$$

where $I(U;(Y,X)^n)$ is the mutual information between the sample U and the sample $(Y,X)^n$. Note now that for the mutual information we have $I(U;(Y,X)^n) = I(U;(X\beta+\varepsilon,X)^n) \leq nH((X\beta+\varepsilon,X)) - nH((X\beta+\varepsilon,X)|U)$*, where the last inequality follows from the chain inequality of entropy.

We therefore need an upper bound on $nH((X\beta+\epsilon,X)) - nH((X\beta+\epsilon,X)|U)$. We can readily calculate $H((X\beta+\epsilon,X)|U)$ as conditioning on the places of the non-zero coordinates of β (WLOG assume they are the first s coordinated, as by a symmetric argument all configurations lead to the same result) we know that the data is normally distributed with covariance matrix given by:

$$A_U = \begin{bmatrix} 1+\sigma^2 & \frac{1}{\sqrt{s}} & \frac{1}{\sqrt{s}} & \dots & \frac{1}{\sqrt{s}} & 0 & \dots & 0 \\ \frac{1}{\sqrt{s}} & 1 & 0 & \dots & 0 & 0 & \dots & 0 \\ \frac{1}{\sqrt{s}} & 0 & 1 & \dots & 0 & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \vdots \\ \frac{1}{\sqrt{s}} & 0 & 0 & \dots & 1 & 0 & \dots & 0 \\ 0 & 0 & 0 & \dots & 0 & 1 & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \dots & 0 & 0 & \dots & 1 \end{bmatrix}.$$

*Here we use H to denote the entropy, not to be confused with the number of slices.

Using the following simple fact:

$$\det \begin{bmatrix} a & v^T \\ v & \mathbb{I} \end{bmatrix} = \det \left(\begin{bmatrix} a & v^T \\ v & \mathbb{I} \end{bmatrix} \begin{bmatrix} 1 & 0 \\ -v & \mathbb{I} \end{bmatrix} \right) = a - v^T v.$$

we have that $|\det A_U| = \sigma^2$. Therefore we have that $H((X\beta + \epsilon, X)|U) = \frac{p+1}{2}(1 + \log(2\pi)) + \frac{1}{2}\log(\sigma^2)$.

We next bound $nH((X\beta + \epsilon, X))$. Note that unconditionally on U, the first coordinate is actually a mixture of normal distributions. Obviously however the data has mean 0, and furthermore, the covariance can be calculated upon noting that $\mathrm{Cov}(X\beta + \epsilon, X) = \mathbb{E}[\mathrm{Cov}(X\beta + \epsilon, X|U)] + \underbrace{\mathrm{Cov}(\mathbb{E}[(X\beta + \epsilon, X)|U])}_{0}$. Thus the covariance is given by:

$$A = \begin{bmatrix} 1+\sigma^2 & \frac{1}{\sqrt{s}} & \frac{1}{\sqrt{s}} & \dots & \frac{1}{\sqrt{s}} & \frac{1}{(p-s+1)\sqrt{s}} & \dots & \frac{1}{(p-s+1)\sqrt{s}} \\ \frac{1}{\sqrt{s}} & 1 & 0 & \dots & 0 & 0 & \dots & 0 \\ \frac{1}{\sqrt{s}} & 0 & 1 & \dots & 0 & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \vdots \\ \frac{1}{\sqrt{s}} & 0 & 0 & \dots & 1 & 0 & \dots & 0 \\ \frac{1}{(p-s+1)\sqrt{s}} & 0 & 0 & \dots & 0 & 1 & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ \frac{1}{(p-s+1)\sqrt{s}} & 0 & 0 & \dots & 0 & 0 & \dots & 1 \end{bmatrix},$$

where the number of $\frac{1}{\sqrt{s}}$ is $s - 1$. Direct evaluation of the determinant yields $|\det(A)| = \sigma^2 + \frac{1}{s} - \frac{1}{s(p-s+1)}$. Therefore by the inequality $\log(1 + x) \leq x$, since the entropy is actually bounded by the entropy of a normal distribution with the same mean and covariance matrix we have that $H((X\beta + \epsilon, X)) \leq \frac{p+1}{2}(1 + \log(2\pi)) + \frac{1}{2}\log(\sigma^2 + \frac{1}{s} - \frac{1}{s(p-s+1)}) \leq \frac{p+1}{2}(1 + \log(2\pi)) +$

$\frac{1}{2}\log(\sigma^2) + \frac{1}{2\sigma^2}\left(\frac{1}{s} - \frac{1}{s(p-s+1)}\right)$. This finally yields the following upper bound on the mutual information:

$$\frac{\frac{n}{s} - \frac{n}{s(p-s+1)}}{2\sigma^2} = \frac{1-C_V}{C_V}\frac{\frac{n}{s} - \frac{n}{s(p-s+1)}}{2}.$$

The above bound is of course $\leq \frac{1-C_V}{C_V}\frac{n}{4s}$, when $p \geq s+1$, which is clearly true in the case when $s = O(p^{1-\delta})$. Therefore if $n < \frac{C_V}{1-C_V}2s\log(p-s+1)$ we will have errors with probability at least $\frac{1}{2}$, asymptotically.

To finish the conclusion, note that the sum (B.1.7), can be split into $\binom{p}{s-1}$ terms, by the following operation:

1. Repeat each set in $\mathbb{S}$ — s times, and denote this superset by $s \times \mathbb{S}$
2. For each S of the $\binom{p}{s-1}$, subsets of $[p]$ with $s-1$ elements, collect $p-s+1$ distinct elements of $s \times \mathbb{S}$ containing S
3. Apply the $\frac{1}{2}$ error bound obtained from above to this local sum.

In the end we get that the probability of error by selecting $S \subset \mathbb{S}$ uniformly is at least: $\frac{1}{s}\frac{\binom{p}{s-1}}{\binom{p}{s}}(p-s+1)\frac{1}{2} = \frac{1}{2}$. □

B.2 Verification of the DT/SDP Constants

B.2.1 DT Constants

In this section we show that the constants defined in (3.5.4), (3.5.5), (3.5.6) and (3.4.13) satisfied the conditions so that the probability of the event $\widetilde{\widetilde{S}}$ goes to 1, and further the 6 terms are bounded by $\frac{C_V}{s}$.

We first start by verifying the probability requirements. Notice that by the definitions of H and ϵ — (3.5.4), (3.5.5) — we have $H > M$, $\frac{1}{H} + 2\epsilon \leq \frac{K}{H}$, $\frac{1}{H} - 2\epsilon \geq \frac{1}{H} - 2\frac{1}{4H} \geq \frac{1}{2H}$.

Let $r = \max(\log(s+1), \log(p-s))$ for brevity. Note that by definition $\eta^2(m-1) \geq 2\widetilde{C}_4(1+\gamma)r$. Recall that $q = \frac{1}{H} - 2\epsilon$. Note that:

$$\begin{aligned}\widetilde{C}_4 &= \widetilde{C}_0 + \widetilde{C}_1 + \widetilde{C}_2\sqrt{-\log\frac{1}{4H}} + \widetilde{C}_3\left(-\log\left(\frac{K}{2H}\right)\frac{K}{H}\right) \\ &\geq \frac{\widetilde{C}_0}{\sqrt{s}} + \widetilde{C}_1 + \widetilde{C}_2\sqrt{-\log\frac{q}{2}} + \widetilde{C}_3\left(-\log\left(\frac{q}{2}\right)q\right),\end{aligned}$$

where we used $\widetilde{C}_0 \geq 1$, $\sqrt{-\log x}$ is decreasing and the function $-\log(x/2)x$ is increasing on $(0, \frac{2}{e})$, but we have $\frac{K}{H} \leq \frac{2}{e}$ by the definition of H (3.5.4). The above bounds clearly give:

$$\begin{aligned}&sH\exp\left(-\frac{1}{2}\frac{\eta^2(m-1)}{\eta + \widetilde{C}_1 + \widetilde{C}_2\sqrt{-\log\left(\frac{q}{2}\right)} + \widetilde{C}_3\left(-\log\left(\frac{q}{2}\right)q\right)}\right) \\ &\leq sH\exp(-(1+\gamma)r).\end{aligned}$$

The above clearly goes to 0, with $p \to \infty$. Since $n = mH$, H is fixed and $m \to \infty$, we have:

$$\exp(-2n\epsilon^2) \leq \exp\left(-2\frac{m}{H}\left(\min\left\{(K-1)/2, \frac{1-l}{2}, 1/54\right\}\right)^2\right) \to 0.$$

Furthermore, $s\exp(-3/16n\tau^2) \to 0$. As $n \gg s$ the above convergence poses no issues. This covers the probability bounds.

Next we deal with showing that each of the 6 terms defined in inequality (3.4.8) and Lemma 3.4.3 is $\leq \frac{C_V}{12s}$. We start with

$$\frac{CH^\kappa}{s}\left(\frac{1}{H} + 2\epsilon\right) \leq \frac{CH^\kappa}{s}\frac{K}{H} \leq \frac{CKH^{\kappa-1}}{s} \leq \frac{C_V}{12s},$$

as promised. We proceed with bounding

$$\begin{aligned}\frac{\frac{C_V}{s}+\frac{CH^\kappa}{s}\left(\frac{1}{H}+2\epsilon\right)}{\left(\frac{1}{H}-2\epsilon\right)} &\leq \frac{1}{\frac{1}{H}-2\epsilon}\left(\frac{C_V}{s}+\frac{C_V}{12s}\right)\\ &\leq 2H\left(\frac{C_V}{s}+\frac{C_V}{12s}\right), \end{aligned} \tag{B.2.1}$$

where the last inequality follows from $\frac{1}{H}-2\epsilon \geq \frac{1}{2H}$. Next we consider the term:

$$\begin{aligned}\left(2\epsilon+\frac{1}{H}-\frac{(m-1)^2}{Hm^2}\right)\frac{\frac{C_V}{s}+\frac{CH^\kappa}{s}\left(\frac{1}{H}+2\epsilon\right)}{\left(\frac{1}{H}-2\epsilon\right)} &\leq \frac{2\epsilon}{\frac{1}{H}-2\epsilon}\left(\frac{C_V}{s}+\frac{C_V}{12s}\right)\\ &\quad+2\left(1-\frac{(m-1)^2}{m^2}\right)\left(\frac{C_V}{s}+\frac{C_V}{12s}\right),\end{aligned}$$

where we used (B.2.1) in the last line. Since $\epsilon \leq \frac{1}{54H}$ we have that $\frac{2\epsilon}{\frac{1}{H}-2\epsilon} \leq \frac{1}{26}$. Thus the first term:

$$\frac{2\epsilon}{\frac{1}{H}-2\epsilon}\left(\frac{C_V}{s}+\frac{C_V}{12s}\right) \leq \frac{C_V}{24s}.$$

For the second term notice that $m \geq 104$ and thus $\left(1-\frac{(m-1)^2}{m^2}\right) \leq \frac{1}{52}$. This and (B.2.1) imply that:

$$2\left(1-\frac{(m-1)^2}{m^2}\right)\left(\frac{C_V}{s}+\frac{C_V}{12s}\right) \leq \frac{C_V}{24s}.$$

This confirms that:

$$\left(2\epsilon+\frac{1}{H}-\frac{(m-1)^2}{Hm^2}\right)\frac{\frac{C_V}{s}+\frac{CH^\kappa}{s}\left(\frac{1}{H}+2\epsilon\right)}{\left(\frac{1}{H}-2\epsilon\right)} \leq \frac{C_V}{12s}.$$

We proceed with the term: $\frac{1+\tau}{m} < \frac{C_V}{12s}$, as we can see that $m > \frac{\widetilde{C}_5}{C_V}s > \frac{12}{C_V}s$. Next notice that

obviously, by the definition of η (3.5.6):

$$\eta^2 \leq \frac{C_V}{12s}. \tag{B.2.2}$$

Next we deal with:

$$\frac{2\sqrt{1+\tau}}{\sqrt{m}}\sqrt{2\eta^2 + 2\frac{1}{H}\frac{\frac{C_V}{s} + \frac{CH^\kappa}{s}\left(\frac{1}{H} + 2\epsilon\right)}{\left(\frac{1}{H} - 2\epsilon\right)}} \leq \frac{2\sqrt{1+\tau}}{\sqrt{m}}\sqrt{\frac{C_V}{6s} + 4\frac{C_V}{s} + \frac{C_V}{3s}}.$$

Where we used (B.2.1) and (B.2.2). Notice now that $\sqrt{\widetilde{C}_5} = 12\left(2\sqrt{1+\tau}\sqrt{\frac{1}{6}+\frac{1}{3}+4}\right)$. This implies:

$$\frac{2\sqrt{1+\tau}}{\sqrt{m}}\sqrt{2\eta^2 + 2\frac{1}{H}\frac{\frac{C_V}{s} + \frac{CH^\kappa}{s}\left(\frac{1}{H} + 2\epsilon\right)}{\left(\frac{1}{H} - 2\epsilon\right)}} \leq \frac{C_V}{12s\sqrt{r}},$$

as recall $m \geq \frac{\widetilde{C}_5}{C_V} sr$. This is even a little smaller than promised.

Finally, we investigate the last term:

$$2\eta\frac{1}{H}\frac{\sqrt{\frac{C_V}{s} + \frac{CH^\kappa}{s}\left(\frac{1}{H} + 2\epsilon\right)}}{\left(\frac{1}{H} - 2\epsilon\right)} \leq 4\eta\sqrt{\left(1 + \frac{1}{12}\right)\frac{C_V}{s}}.$$

Note that $\eta = \frac{1}{48\sqrt{1+\frac{1}{12}}}\frac{\sqrt{C_V}}{\sqrt{s}}$, which gives:

$$2\eta\frac{1}{H}\frac{\sqrt{\frac{C_V}{s} + \frac{CH^\kappa}{s}\left(\frac{1}{H} + 2\epsilon\right)}}{\left(\frac{1}{H} - 2\epsilon\right)} \leq \frac{C_V}{12s}.$$

B.2.2 SDP Constants

We start by verifying that the probability bounds converge to 0, so that asymptotically the support is recovered with probability 1. We begin with:

$$p_1 = sH \exp\left(-\frac{1}{2}\frac{\eta^2(m-1)}{\eta + \widetilde{C}_1 + \widetilde{C}_2\sqrt{-\log\left(\frac{q}{2}\right)} + \widetilde{C}_3\left(-\log\left(\frac{q}{2}\right)q\right)}\right).$$

Recall that $q = \frac{1}{H} - 2\epsilon$. Just as before we have the bound:

$$\begin{aligned}&\widetilde{C}_0' + \widetilde{C}_1 + \widetilde{C}_2\sqrt{\log\frac{2H}{l}} + \widetilde{C}_3\left(\log\left(\frac{2H}{K}\right)\frac{K}{H}\right)\\ &\geq \eta + \widetilde{C}_1 + \widetilde{C}_2\sqrt{-\log\frac{q}{2}} + \widetilde{C}_3\left(-\log\left(\frac{q}{2}\right)q\right).\end{aligned}$$

where we used $\widetilde{C}_0' = \frac{l\sqrt{C_V}}{48} > \eta$, $\sqrt{-\log x}$ is decreasing and the function $-\log(x/2)x$ is increasing on $(0, \frac{2}{e})$, but we have $\frac{K}{H} \leq \frac{2}{e}$ by the definition of H. Furthermore note that: $\log(x) \leq x$, and thus $\left(\log\left(\frac{2H}{K}\right)\frac{K}{H}\right) \leq 2$. Moreover, since $\sqrt{a+b} \leq \sqrt{a} + \sqrt{b}$, we have $\sqrt{\log(2H/l)} \leq \sqrt{\log(2/l)} + \sqrt{\log(H)}$. Define the constant $\widetilde{C}' = \widetilde{C}_0' + \widetilde{C}_1 + \widetilde{C}_2\sqrt{\log(2/l)} + 2\widetilde{C}_3$, and we have:

$$\widetilde{C}' + \widetilde{C}_2\sqrt{\log(H)} \geq \eta + \widetilde{C}_1 + \widetilde{C}_2\sqrt{-\log\frac{q}{2}} + \widetilde{C}_3\left(-\log\left(\frac{q}{2}\right)q\right).$$

Note that by the definition of m, the following holds:

$$m \geq s\frac{16(12t+1)^2(\log(s)+\log(H))(\widetilde{C}' + \widetilde{C}_2\sqrt{\log(H)})}{l^2 C_V}t.$$

Thus:

$$\frac{m}{s} - \frac{16(12t+1)^2(\log(s)+\log(H))(\widetilde{C}' + \widetilde{C}_2\sqrt{\log(H)})}{l^2 C_V} \to \infty,$$

which readily implies that $p_1 \to 0$.

Next we need to control:

$$p_2 = \exp(-2n\epsilon^2) \leq \exp\left(-2\frac{m}{H}\left(\min\left\{(K-1)/2, \frac{1-l}{2}, \frac{l}{4(1+48t)}\right\}\right)^2\right).$$

Since we have selected: $m \geq Ht^3$, we have that $p_2 \to 0$. Finally, it's clear that $s\exp(-3/16n\tau^2) \to 0$, as $n \gg s$.

Next we turn our attention to showing that each of the 6 terms in the bound (3.5.3) is $\leq \frac{C_V}{12st}$. Before we proceed we observe several useful inequalities which follow immediately from the definitions:

$$\frac{1}{H} - 2\epsilon \geq \frac{l}{H}, \frac{1}{H} + 2\epsilon \leq \frac{K}{H}$$

Moreover, obviously we have that $t \geq 1$.

Since: $H \geq \left(\frac{12CKt}{C_V}\right)^{\frac{1}{1-\kappa}}$, we have:

$$\frac{CH^\kappa}{s}\left(\frac{1}{H} + 2\epsilon\right) \leq \frac{CH^\kappa}{s}\frac{K}{H} \leq \frac{CKH^{\kappa-1}}{s} \leq \frac{C_V}{12st}.$$

Next, since $\epsilon \leq \frac{l}{4H(1+12t)}$ we have:

$$\begin{aligned}
2\epsilon\frac{\frac{C_V}{s} + \frac{CH^\kappa}{s}\left(\frac{1}{H} + 2\epsilon\right)}{\left(\frac{1}{H} - 2\epsilon\right)} &\leq 2\epsilon\frac{H}{l}\left(\frac{C_V}{s} + \frac{CH^\kappa}{s}\left(\frac{1}{H} + 2\epsilon\right)\right) \\
&\leq 2\epsilon\frac{H}{l}\left(\frac{C_V}{s} + \frac{C_V}{12st}\right) \\
&\leq \frac{C_V}{24st}.
\end{aligned}$$

Next, using the fact that $m \geq 4(12t+1)\frac{1}{l}$ we have:

$$\begin{aligned}\frac{1}{H}\left(1-\frac{(m-1)^2}{m^2}\right)\frac{\frac{C_V}{s}+\frac{CH^\kappa}{s}\left(\frac{1}{H}+2\epsilon\right)}{\left(\frac{1}{H}-2\epsilon\right)} &\leq \frac{2}{ml}\left(\frac{C_V}{s}+\frac{C_V}{12st}\right)\\ &\leq \frac{C_V}{24st}.\end{aligned}$$

Since $m \geq \frac{(1+\tau)48st}{C_V}$, we have: $4\frac{1+\tau}{m} \leq \frac{C_V}{12st}$.

Note that it follows from the definition of η that $\eta \leq \sqrt{\frac{C_V}{48st}}$, and hence:

$$4\eta^2 \leq \frac{C_V}{12st}$$

Next, consider:

$$\begin{aligned}\frac{4\sqrt{1+\tau}}{\sqrt{m}}\sqrt{2\eta^2+2\frac{1}{H}\frac{\frac{C_V}{s}+\frac{CH^\kappa}{s}\left(\frac{1}{H}+2\epsilon\right)}{\left(\frac{1}{H}-2\epsilon\right)}} &\leq \frac{4\sqrt{1+\tau}}{\sqrt{m}}\sqrt{\frac{C_V}{24st}+\frac{2}{l}\frac{C_V}{s}+\frac{1}{l}\frac{C_V}{6st}}\\ &\leq \frac{C_V}{12st}.\end{aligned}$$

where the last inequality holds since:

$$m \geq (1+\tau)48^2st^2\frac{\left(\frac{1}{24t}+\frac{2}{l}+\frac{1}{l6t}\right)}{C_V},$$

by the definition of m, and since $t \geq 1$.

Finally, we investigate the last term:

$$\begin{aligned}4\eta\frac{1}{H}\frac{\sqrt{\frac{C_V}{s}+\frac{CH^\kappa}{s}\left(\frac{1}{H}+2\epsilon\right)}}{\left(\frac{1}{H}-2\epsilon\right)} &\leq \frac{4\eta}{l}\sqrt{\frac{C_V}{s}+\frac{C_V}{12st}}\\ &\leq \frac{C_V}{12st},\end{aligned}$$

where the last inequality holds since:

$$\eta \leq \frac{l}{4} \frac{\sqrt{C_V}}{\sqrt{12t}\sqrt{12t+1}\sqrt{s}},$$

by definition. With this the verification of the constants is complete.

B.3 Collection of Useful Lemmas

In this section, for convenience of the reader, we restate several lemmas that we use often in our analysis.

Lemma B.3.1 (Lemma 5[87]). *Consider a fixed nonzero vector $z \in \mathbb{R}^s$ and a random matrix $A_{n\times s}$, whose entries are iid standard normal random variables. There are positive constants C_1 and C_2 such that for all $t > 0$:*

$$\mathbb{P}\left(||[(n^{-1}A^\intercal A) - \mathbb{I}_{s\times s}]z||_\infty \geq C_1||z||_\infty\right) \leq 4\exp(-C_2\min(s, \log(p-s))),$$

where $C_1 = \sqrt{\frac{s\log(p-s)}{C_3 n}}$, for some absolute constant $C_3 > 0$.

Lemma B.3.2 (Corollary 5.35[84]). *Let $A_{n\times s}$ matrix whose entries are iid standard normal random variables. Then for every $t \geq 0$, with probability at least $1 - 2\exp(-t^2/2)$ one has:*

$$\sqrt{n} - \sqrt{s} - t \leq s_{\min}(A) \leq s_{\max}(A) \leq \sqrt{n} + \sqrt{s} + t,$$

where $s_{\min}(A)$ and $s_{\max}(A)$ are the smallest and largest singular values of A correspondingly.

B.4 Covariance Thresholding

Proof of Lemma 3.6.7. Note that, exponential concentration bounds do not apply in this case. However, observe that by the properties of the multivariate normal distribution projecting X in the space $\beta^{\perp}$ by $(\mathbb{I} - \beta\beta^{\intercal})X$ makes it independent of Y. Clearly, the random variable $Y(\mathbb{I} - \beta\beta^{\intercal})X$ has mean 0. Note that conditionally on $Y_i, i = 1, \ldots, n$ for any $j \in \{1, \ldots, p\}$ we have that $\frac{1}{n}\sum_{i=1}^{n} Y_i[(\mathbb{I} - \beta\beta^{\intercal})X_i]^j \sim N(0, n^{-2}\sum_{i=1}^{n} Y_i^2 \left[(\mathbb{I} - \beta\beta^{\intercal})\right]_{jj})$. Clearly we have,

$$n^{-2}\sum_{i=1}^{n} Y_i^2 \left[(\mathbb{I} - \beta\beta^{\intercal})\right]_{jj} \leq n^{-2}\sum_{i=1}^{n} Y_i^2,$$

for all j. Thus by a standard Gaussian tail bound:

$$\mathbb{P}\left(\left|\left|\frac{1}{n}\sum_{i=1}^{n} Y_i(\mathbb{I} - \beta\beta^{\intercal})X_i\right|\right|_{\infty} \geq t \middle| \mathbf{Y}\right) \leq 2p\exp\left[-\frac{nt^2}{2\overline{Y^2}}\right],$$

where $\overline{Y^2} = n^{-1}\sum_{i=1}^{n} Y_i^2$. By Chebyshev's inequality $\mathbb{P}(|\overline{Y^2} - \sigma^2| \geq r) \leq \frac{\eta}{nr^2}$. Hence selecting $r = \sqrt{\frac{\log n}{n}}$ will keep the above probability going to 1 at rate $\frac{\eta}{\log n}$ and moreover for large n we have $\overline{Y^2} \leq \sigma^2 + 1$. Using this bound in the tail bound above yields that for a choice of $t = 2\sqrt{(\sigma^2+1)\frac{\log p}{n}}$ the tail bound will go to 0 at a rate $\frac{2}{p}$, as claimed.

Next consider controlling:

$$\mathbb{P}\left(\left|\left|\frac{1}{n}\sum_{i=1}^{n} Y_i\beta\beta^{\intercal}X_i - c_0\beta\right|\right|_{\infty} \geq t\right) = \mathbb{P}\left(\left|\frac{1}{n}\sum_{i=1}^{n} Y_i\beta^{\intercal}X_i - c_0\right| \geq t/||\beta||_{\infty}\right),$$

where $E[YX] = c_0\beta$, and c_0 is defined in the main text. Applying Chebyshev's inequality once again we get that $t = \frac{||\beta||_{\infty}\sqrt{\log n}}{\sqrt{n}}$ suffices to keep the above probability going to 0. By the triangle

inequality we conclude that, with probability going to 1:

$$\left|\left|\frac{1}{n}\sum_{i=1}^{n} Y_i X_i - E[YX]\right|\right|_\infty \leq \frac{||\beta||_\infty \sqrt{\log n}}{\sqrt{n}} + 2\sqrt{(\sigma^2+1)\frac{\log p}{n}}.$$

This is what we claimed. □

B.5 LASSO Support Recovery

Proof of Lemma 3.6.13. Note that since $P_{\mathbf{X}_S^\perp}$ is an orthogonal projection matrix it contracts length and hence:

$$\left|\left|P_{\mathbf{X}_S^\perp}\left(\frac{\mathbf{w}}{\lambda n}\right)\right|\right|_2^2 \leq \frac{||\mathbf{w}||_2^2}{\lambda^2 n^2}.$$

Next observe that $\mathbf{w} = \mathbf{Y} - c_0\mathbf{X}\beta^*$ is a vector with non-zero mean. However, by Chebyshev's inequality we have:

$$\mathbb{P}\left(\left|\frac{||\mathbf{w}||_2^2}{n} - \xi^2\right| \geq t\right) \leq \frac{\theta^2}{nt^2}.$$

Then setting $t = 1$ brings the above probability to 0 at a rate $\frac{\theta^2}{n}$. Next:

$$n^{-1}\check{z}_S^\intercal (n^{-1}\mathbf{X}_S^\intercal \mathbf{X}_S)^{-1}\check{z}_S \leq \frac{1}{\lambda_{\min}^S(1-2\sqrt{\frac{s}{n}})^2}\frac{||\check{z}_S||_2^2}{n} \leq \frac{1}{\lambda_{\min}^S(1-2\sqrt{\frac{s}{n}})^2}\frac{s}{n},$$

with probability at least $1 - 2\exp(-s/2)$, where we used Lemma B.3.2. This completes the proof.

□

Proof of Lemma 3.6.15. First, we note the following decomposition:

$$[\mathbf{X}^\intercal\mathbf{X}]^{-1}\mathbf{X}^\intercal\mathbf{Y} - c_0\beta^* = (n[\mathbf{X}^\intercal\mathbf{X}]^{-1} - \mathbb{I})n^{-1}\mathbf{X}^\intercal\mathbf{Y} + (n^{-1}\mathbf{X}^\intercal\mathbf{Y} - c_0\beta^*).$$

Note that the second term is mean 0. Applying Lemma 3.6.7 gives us a bound on the second term.

We next move on to consider the first term.

Consider a "symmetrization" transformation of the predictor matrix $\widetilde{\mathbf{X}}^\intercal = (\mathbb{I} - \beta^*\beta^{*\intercal})\mathbf{X}^\intercal + \beta^*\beta^{*\intercal}\mathbf{X}^{*\intercal}$, where $\mathbf{X}^*_{n\times s}$ is an iid copy of $\mathbf{X}$, or in other words the columns of $\mathbf{X}^*$: $X_i^* \sim N(0, \mathbb{I}_{n\times n}), i = 1, \ldots, s$ and are independent of $\mathbf{X}$. Note that in doing this construction, we guarantee that $\widetilde{\mathbf{X}}$ is independent of $\mathbf{X}^\intercal\beta^*$. We will need the following result:

Lemma B.5.1. *Suppose that s, n satisfy $\frac{s}{n} \leq \frac{1}{64}$. The following bound holds:*

$$||[n^{-1}\mathbf{X}^\intercal\mathbf{X}]^{-1} - [n^{-1}\widetilde{\mathbf{X}}^\intercal\widetilde{\mathbf{X}}]^{-1}||_{\infty,\infty} \leq 8\sqrt{s}\left(C_1||\beta^*||_\infty + 4\sqrt{\frac{\log s}{n}}\right),$$

with probability at least $1 - 8\exp(-C_2\min(s, \log(p-s))) - 4\exp(-s/2) - \frac{4}{s} - \frac{4}{n}$, where $C_1 > 0$ and $C_2 > 0$ are constants.

Remark B.5.2. *The constant $C_1 = \sqrt{\frac{256s\log(p-s)}{C_3 n}}$ with $C_3 > 0$ being an absolute constant, can be chosen to be arbitrary small, for the sake of making n proportionally large comparable to $s\log(p-s)$.*

Now we further decompose the first term as follows:

$$\begin{aligned}(n[\mathbf{X}^\intercal\mathbf{X}]^{-1} - \mathbb{I})n^{-1}\mathbf{X}^\intercal\mathbf{Y} = (n[\mathbf{X}^\intercal\mathbf{X}]^{-1} - \mathbb{I})\beta^*\beta^{*\intercal}n^{-1}\mathbf{X}^\intercal\mathbf{Y} \\ + n([\mathbf{X}^\intercal\mathbf{X}]^{-1} - [\widetilde{\mathbf{X}}^\intercal\widetilde{\mathbf{X}}]^{-1})(\mathbb{I} - \beta^*\beta^{*\intercal})n^{-1}\mathbf{X}^\intercal\mathbf{Y} \\ + (n[\widetilde{\mathbf{X}}^\intercal\widetilde{\mathbf{X}}]^{-1} - \mathbb{I})n^{-1}\widetilde{\mathbf{X}}^\intercal\mathbf{Y} \\ - (n[\widetilde{\mathbf{X}}^\intercal\widetilde{\mathbf{X}}]^{-1} - \mathbb{I})\beta^*\beta^{*\intercal}n^{-1}\mathbf{X}^{*\intercal}\mathbf{Y}.\end{aligned}$$

We next deal with each of these terms separately. For the first and last term we can apply Lemma B.3.1. Under the same event as in Lemma B.5.1 we have that $||([n^{-1}\mathbf{X}^\intercal\mathbf{X}]^{-1} - \mathbb{I})\beta^*||_\infty \leq C_1||\beta^*||_\infty$ and $||([n^{-1}\widetilde{\mathbf{X}}^\intercal\widetilde{\mathbf{X}}]^{-1} - \mathbb{I})\beta^*||_\infty \leq C_1||\beta^*||_\infty$. Furthermore, $\beta^{*\intercal}\mathbf{X}^\intercal\mathbf{Y}/n$ is a mean c_0 random variable. Just as in the proof of Lemma 3.6.7 by Chebyshev's inequality we have that with probability at

least $1 - \frac{\gamma}{\log n}$ we have $|\beta^\intercal \mathbf{X}^\intercal \mathbf{Y}/n| \leq |c_0| + \sqrt{\frac{\log n}{n}}$.

Furthermore, notice that $n^{-1}\beta^{*\intercal}\mathbf{X}^{*\intercal}\mathbf{Y}$ is a mean 0 random variable. Conditionally on $\mathbf{Y}$ it has a $N(0, n^{-2}\sum Y_i^2)$ distribution. With exactly the same argument as in the proof of Lemma 3.6.7 we conclude that with probability at least $1 - \frac{\eta}{\log n} - \frac{2}{n}$:

$$|n^{-1}\beta^{*\intercal}\mathbf{X}^{*\intercal}\mathbf{Y}| \leq 2\sqrt{(\sigma^2+1)\frac{\log n}{n}}.$$

Thus, combining these results we get:

$$\begin{aligned}||(n[\mathbf{X}^\intercal\mathbf{X}]^{-1} - \mathbb{I})n^{-1}\mathbf{X}^\intercal\mathbf{Y}||_\infty \leq C_1||\beta^*||_\infty &\left(|c_0| + \sqrt{\frac{\log n}{n}} + 2\sqrt{(\sigma^2+1)\frac{\log n}{n}}\right) \\ &+ ||n([\mathbf{X}^\intercal\mathbf{X}]^{-1} - [\widetilde{\mathbf{X}}^\intercal\widetilde{\mathbf{X}}]^{-1})(\mathbb{I} - \beta^*\beta^{*\intercal})n^{-1}\mathbf{X}^{*\intercal}\mathbf{Y}||_\infty \\ &+ ||(n[\widetilde{\mathbf{X}}^\intercal\widetilde{\mathbf{X}}]^{-1} - \mathbb{I})n^{-1}\widetilde{\mathbf{X}}^\intercal\mathbf{Y}||_\infty.\end{aligned}$$

To deal with the second first note that:

$$\begin{aligned}&||n([\mathbf{X}^\intercal\mathbf{X}]^{-1} - [\widetilde{\mathbf{X}}^\intercal\widetilde{\mathbf{X}}]^{-1})(\mathbb{I} - \beta^*\beta^{*\intercal})n^{-1}\mathbf{X}^\intercal\mathbf{Y}||_\infty \\ &\leq ||n([\mathbf{X}^\intercal\mathbf{X}]^{-1} - [\widetilde{\mathbf{X}}^\intercal\widetilde{\mathbf{X}}]^{-1})||_{\infty,\infty}||(\mathbb{I} - \beta^*\beta^{*\intercal})n^{-1}\mathbf{X}^\intercal\mathbf{Y}||_\infty.\end{aligned}$$

Note that the second term is a mean 0 random variable since $(\mathbb{I} - \beta^*\beta^{*\intercal})\mathbf{X}$ is independent of $\mathbf{Y}$. In Lemma 3.6.7 we argued that $||(\mathbb{I} - \beta^*\beta^{*\intercal})n^{-1}\mathbf{X}^\intercal\mathbf{Y}||_\infty \leq 2\sqrt{(\sigma^2+1)\frac{\log s}{n}}$ with probability at least $1 - \frac{2}{s}$ (this event is in fact a sub-event of the bounds of the first term $n^{-1}\mathbf{X}^\intercal\mathbf{Y} - c_0\beta^*$).

By Lemma B.5.1 we have that the term $||n([\mathbf{X}^\intercal\mathbf{X}]^{-1} - [\widetilde{\mathbf{X}}^\intercal\widetilde{\mathbf{X}}]^{-1})||_{\infty,\infty} = O(1)$, which covers the third term. Note here that by $O(1)$ we mean that this term can be bounded by arbitrarily small constant with high probability at the expense for making the ratio $\frac{n}{s\log(p-s)}$ high.

Finally to deal with the last term we will make use of the following:

Lemma B.5.3. *Let* $\frac{s}{n} \leq \frac{1}{64}$. *Then there exists a constant* $\Omega \asymp \sigma > 0$, *such that the term:*

$$||(n[\widetilde{\mathbf{X}}^\intercal \widetilde{\mathbf{X}}]^{-1} - \mathbb{I})n^{-1}\widetilde{\mathbf{X}}^\intercal \mathbf{Y}||_\infty \leq \Omega\sqrt{\frac{\log s}{n}},$$

with probability at least $1 - \frac{2}{s} - \frac{\eta}{\log n} - 2\exp(-s/2)$.

Applying Lemma B.5.3 we have in conjunction with our previous bounds we get:

$$\begin{aligned}||[\mathbf{X}^\intercal \mathbf{X}]^{-1}\mathbf{X}^\intercal \mathbf{Y} - c_0\beta^*||_\infty \leq C_1||\beta^*||_\infty &\left(|c_0| + \sqrt{\frac{\log n}{n}} + 2\sqrt{(\sigma^2+1)\frac{\log n}{n}}\right) \\ &+ 16\sqrt{s}\left(C_1||\beta^*||_\infty + 4\sqrt{\frac{\log s}{n}}\right)\sqrt{(\sigma^2+1)\frac{\log s}{n}} \\ &+ \Omega\sqrt{\frac{\log s}{n}} + \frac{||\beta^*||_\infty\sqrt{\log n}}{\sqrt{n}} + 2\sqrt{(\sigma^2+1)\frac{\log s}{n}},\end{aligned}$$

with probability at least $1 - 8\exp(-C_2\min(s, \log(p-s))) - 4\exp(-s/2) - \frac{6}{n} - \frac{8}{s} - 2\frac{\eta+\gamma}{\log n}$[†], which finishes the proof, after grouping terms. □

Proof of Lemma B.5.1. We first compare $[n^{-1}\mathbf{X}^\intercal\mathbf{X}]^{-1}$ to $[n^{-1}\widetilde{\mathbf{X}}^\intercal\mathbf{X} + \beta^*\beta^{*\intercal}]^{-1}$. Observe that the latter matrix is invertible wap[‡] 1. Note that $n^{-1}\widetilde{\mathbf{X}}^\intercal\mathbf{X} + \beta^*\beta^{*\intercal} = (\mathbb{I} - \beta^*\beta^{*\intercal})n^{-1}\mathbf{X}^\intercal\mathbf{X} + \beta^*\beta^{*\intercal}(\mathbb{I} + n^{-1}\mathbf{X}^{*\intercal}\mathbf{X})$, and since the matrix $\mathbf{X}^\intercal\mathbf{X}$ is full rank wap 1 and $\beta^{*\intercal}(\mathbb{I} + n^{-1}\mathbf{X}^{*\intercal}\mathbf{X}) \neq 0$ wap 1, the matrix in question is of full column rank. Using Woodbury's matrix identity we have:

$$[n^{-1}\widetilde{\mathbf{X}}^\intercal\mathbf{X} + \beta^*\beta^{*\intercal}]^{-1} - [n^{-1}\mathbf{X}^\intercal\mathbf{X}]^{-1} = \frac{[n^{-1}\mathbf{X}^\intercal\mathbf{X}]^{-1}\beta^*\beta^{*\intercal}M[n^{-1}\mathbf{X}^\intercal\mathbf{X}]^{-1}}{1 - \beta^{*\intercal}M[n^{-1}\mathbf{X}^\intercal\mathbf{X}]^{-1}\beta^*},$$

where $M = n^{-1}\mathbf{X}^\intercal\mathbf{X} - \mathbb{I} - n^{-1}\mathbf{X}^{*\intercal}\mathbf{X}$. Next we handle the term $\beta^{*\intercal}M[n^{-1}\mathbf{X}^\intercal\mathbf{X}]^{-1}$. By the

[†]Here we are recognizing the fact that the events of some probability bounds we derived above, in fact coincide.

[‡]Here and throughout wap means "with asymptotic probability".

triangle inequality have:

$$||\beta^{*\intercal} M[n^{-1}\mathbf{X}^\intercal\mathbf{X}]^{-1}||_\infty \leq ||\beta^{*\intercal}([n^{-1}\mathbf{X}^\intercal\mathbf{X}]^{-1} - \mathbb{I})||_\infty + ||\beta^{*\intercal} n^{-1}\mathbf{X}^{*\intercal}\mathbf{X}[n^{-1}\mathbf{X}^\intercal\mathbf{X}]^{-1}||_\infty.$$

For the first term Lemma B.3.1 is directly applicable. Applying this lemma gives us the existence of constants C_1 and C_2 such that $||([n^{-1}\mathbf{X}^\intercal\mathbf{X}]^{-1} - \mathbb{I})\beta^*||_\infty \leq C_1||\beta^*||_\infty$ with probability at least $1 - 4\exp(-C_2\min(s, \log(p-s)))$. For the second term, we have that conditionally on $\mathbf{X}$ it has a normal distribution: $N(0, n^{-1}(n^{-1}\mathbf{X}^\intercal\mathbf{X})^{-1})$. Since $\mathbf{X}$ is standard normal, we can apply Lemma B.3.2 to claim that $||n(\mathbf{X}^\intercal\mathbf{X})^{-1}||_2 \leq \left(\frac{1}{1-\sqrt{\frac{s}{n}}-t}\right)^2$ with probability at least $1 - 2\exp(-nt^2/2)$. Taking $t = \sqrt{\frac{s}{n}}$ gives us that $||n(\mathbf{X}^\intercal\mathbf{X})^{-1}||_2 \leq \frac{1}{(1-2\sqrt{\frac{s}{n}})^2}$ with probability at least $1 - 2\exp(-s/2)$. Thus conditioning on this event, by a standard normal tail bound we have:

$$\mathbb{P}(||\beta^{*\intercal} n^{-1}\mathbf{X}^{*\intercal}\mathbf{X}[n^{-1}\mathbf{X}^\intercal\mathbf{X}]^{-1}||_\infty \geq t) \leq 2s\exp\left(-t^2 n\left(1 - 2\sqrt{\frac{s}{n}}\right)^2/2\right).$$

Selecting $t = 4\sqrt{\frac{\log s}{n}}$, we get the probability above is bounded by $\frac{2}{s}$ (using $\sqrt{\frac{s}{n}} \leq \frac{1}{4}$). So finally on the intersection event we have:

$$||\beta^{*\intercal} M[n^{-1}\mathbf{X}^\intercal\mathbf{X}]^{-1}||_\infty \leq C_1||\beta^*||_\infty + 4\sqrt{\frac{\log s}{n}},$$

with probability at least $1 - \frac{2}{s} - 2\exp(-s/2) - 4\exp(-C_2\min(s, \log(p-s)))$. Let us now consider the denominator:

$$\begin{aligned} 1 - \beta^{*\intercal} M[n^{-1}\mathbf{X}^\intercal\mathbf{X}]^{-1}\beta^* &= 1 - \beta^{*\intercal}(\mathbb{I} - [n^{-1}\mathbf{X}^\intercal\mathbf{X}]^{-1})\beta^* + n^{-1}\beta^{*\intercal}\mathbf{X}^{*\intercal}\mathbf{X}[n^{-1}\mathbf{X}^\intercal\mathbf{X}]^{-1}\beta^* \\ &= \beta^{*\intercal}[n^{-1}\mathbf{X}^\intercal\mathbf{X}]^{-1}\beta^* + n^{-1}\beta^{*\intercal}\mathbf{X}^{*\intercal}\mathbf{X}[n^{-1}\mathbf{X}^\intercal\mathbf{X}]^{-1}\beta^*. \end{aligned}$$

Using Lemma B.3.2 we have $||[n^{-1}\mathbf{X}^\intercal\mathbf{X}]^{-1}||_2 \geq \frac{1}{(1+2\sqrt{\frac{s}{n}})^2} \geq \frac{16}{25}$ with the last bound holding un-

der the condition $\frac{s}{n} \leq \frac{1}{64}$. Hence $|\beta^{*\intercal}[n^{-1}\mathbf{X}^\intercal\mathbf{X}]^{-1}\beta^*| \geq \frac{16}{25}$. For the second term just as before, conditionally on $\mathbf{X}$ we have $n^{-1}\beta^{*\intercal}\mathbf{X}^{*\intercal}\mathbf{X}[n^{-1}\mathbf{X}^\intercal\mathbf{X}]^{-1}\beta^* \sim N(0, n^{-1}\beta^{*\intercal}[n^{-1}\mathbf{X}^\intercal\mathbf{X}]^{-1}\beta^*)$. Then by a standard tail bound we have that the second term is $\leq 4\sqrt{\frac{\log n}{n}}$ with probability at least $1 - \frac{2}{n}$. Putting everything together we have:

$$1 - \beta^{*\intercal}M[n^{-1}\mathbf{X}^\intercal\mathbf{X}]^{-1}\beta^* \geq \frac{16}{25} - 4\sqrt{\frac{\log n}{n}}.$$

The last expression is clearly bigger than $\frac{1}{2}$ for large enough values of n. Hence we conclude that with high probability we have:

$$||[n^{-1}\widetilde{\mathbf{X}}^\intercal\mathbf{X} + \beta^*\beta^{*\intercal}]^{-1} - [n^{-1}\mathbf{X}^\intercal\mathbf{X}]^{-1}||_{\infty,\infty} \leq 2||[n^{-1}\mathbf{X}^\intercal\mathbf{X}]^{-1}\beta||_1||\beta^{*\intercal}M[n^{-1}\mathbf{X}^\intercal\mathbf{X}]^{-1}||_\infty$$

For the first term, by the definition of matrix $||\cdot||_2$ norm we further have:

$$\begin{aligned}||[n^{-1}\mathbf{X}^\intercal\mathbf{X}]^{-1}\beta^*||_1 &\leq \sqrt{s}||[n^{-1}\mathbf{X}^\intercal\mathbf{X}]^{-1}\beta^*||_2 \leq \sqrt{s}||\beta^*||_2||[n^{-1}\mathbf{X}^\intercal\mathbf{X}]^{-1}||_2 \\ &\leq \sqrt{s}\frac{1}{(1 - 2\sqrt{\frac{s}{n}})^2}.\end{aligned}$$

Combining this inequality with our previous bound we get:

$$||[n^{-1}\widetilde{\mathbf{X}}^\intercal\mathbf{X} + \beta^*\beta^{*\intercal}]^{-1} - [n^{-1}\mathbf{X}^\intercal\mathbf{X}]^{-1}||_{\infty,\infty} \leq \frac{2\sqrt{s}}{(1 - 2\sqrt{\frac{s}{n}})^2}\left(C_1||\beta^*||_\infty + 4\sqrt{\frac{\log s}{n}}\right).$$

Next we show that $[n^{-1}\widetilde{\mathbf{X}}^\intercal\mathbf{X} + \beta^*\beta^{*\intercal}]^{-1}$ is also close to $[n^{-1}\widetilde{\mathbf{X}}^\intercal\widetilde{\mathbf{X}}]^{-1}$. By Woodbury's matrix identity we have:

$$[n^{-1}\widetilde{\mathbf{X}}^\intercal\mathbf{X} + \beta^*\beta^{*\intercal}]^{-1} - [n^{-1}\widetilde{\mathbf{X}}^\intercal\widetilde{\mathbf{X}}]^{-1} = \frac{[n^{-1}\widetilde{\mathbf{X}}^\intercal\widetilde{\mathbf{X}}]^{-1}\widetilde{M}\beta^*\beta^{*\intercal}[n^{-1}\widetilde{\mathbf{X}}^\intercal\widetilde{\mathbf{X}}]^{-1}}{1 - \beta^{*\intercal}[n^{-1}\widetilde{\mathbf{X}}^\intercal\widetilde{\mathbf{X}}]^{-1}\widetilde{M}\beta^*},$$

where $\widetilde{M} = n^{-1}\widetilde{\mathbf{X}}^\intercal\widetilde{\mathbf{X}} - \mathbb{I} - n^{-1}\widetilde{\mathbf{X}}^\intercal\mathbf{X}$. Note that since $\widetilde{\mathbf{X}}^\intercal \perp\!\!\!\perp \mathbf{X}\beta^*$, the same argument as before goes through. Combining the bounds with a triangle inequality completes the proof, using the fact that $\sqrt{\frac{s}{n}} \leq \frac{1}{8}$. □

Proof of Lemma B.5.3. We first perform an SVD on the $\widetilde{\mathbf{X}} = U_{n\times s}D_{s\times s}V^\intercal_{s\times s}$ matrix. Noting that since multiplying $\widetilde{\mathbf{X}}$ by a unitary $s \times s$ matrix on the right or with a unitary $n \times n$ matrix on the left doesn't change the distribution of $\widetilde{\mathbf{X}}$, we conclude that the matrices U, D and V are independent. This representation gives us that $(n^{-1}\widetilde{\mathbf{X}}^\intercal\widetilde{\mathbf{X}})^{-1} - \mathbb{I} = V(nD^{-2} - \mathbb{I})V^\intercal$. With this notation we can rewrite:

$$(n[\widetilde{\mathbf{X}}^\intercal\widetilde{\mathbf{X}}]^{-1} - \mathbb{I})n^{-1}\widetilde{\mathbf{X}}^\intercal\mathbf{Y} = V\underbrace{(nD^{-2} - \mathbb{I})n^{-1/2}D}_{W}n^{-1/2}U^\intercal\mathbf{Y}.$$

We recall that by construction $\widetilde{\mathbf{X}}$ is independent of $\mathbf{Y}$. The elements of the matrix W can be bounded in a simple manner. We have $||W||_2 \leq ||(nD^{-2} - \mathbb{I})||_2||n^{-1/2}D||_2$, and by Lemma B.3.2, as before we have: $||(nD^{-2} - \mathbb{I})||_2 \leq \frac{1}{(1-2\sqrt{\frac{s}{n}})^2} - 1 \leq \frac{4\sqrt{\frac{s}{n}}}{(1-2\sqrt{\frac{s}{n}})^2}$ and $||n^{-1/2}D||_2 \leq 1 + 2\sqrt{\frac{s}{n}}$ with probability at least $1 - 2\exp(-s/2)$. We will condition on the event $||W||_2 \leq \frac{4\sqrt{\frac{s}{n}}}{(1-2\sqrt{\frac{s}{n}})^2}(1 + 2\sqrt{\frac{s}{n}}) \leq 9\sqrt{\frac{s}{n}}$, with the last inequality holding for $\sqrt{\frac{s}{n}} \leq \frac{1}{8}$. Since every random variable in the display is independent from W the distributions of V, U and $\mathbf{Y}$ stay unchanged under this conditioning. Let e_i be a unit vector with 1 on the i^{th} position. Since we are interested in bounding the $||\cdot||_\infty$ we will start with the following:

$$e_i^T(n[\widetilde{\mathbf{X}}^\intercal\widetilde{\mathbf{X}}]^{-1} - \mathbb{I})n^{-1}\widetilde{\mathbf{X}}^\intercal\mathbf{Y} = v_i^\intercal W[n^{-1/2}U^\intercal\mathbf{Y}],$$

where $v_i^\intercal$ is the i^{th} row of the matrix V. Condition on the vectors $n^{-1/2}U^\intercal\mathbf{Y}$ and $\mathbf{Y}$. Since v_i is independent of $n^{-1/2}U^\intercal\mathbf{Y}$, $\mathbf{Y}$ it follows that the distribution of v_i is uniform on the unit sphere in

$\mathbb{R}^s$. We next show that the function $F(v_i) = v_i^\intercal W[n^{-1/2}U^\intercal \mathbf{Y}]$ is Lipschitz. We have:

$$\begin{aligned}
|F(v_i) - F(v_i')| &\leq ||v_i - v_i'||_2 ||W||_2 ||n^{-1/2}U^\intercal \mathbf{Y}||_2 \\
&\leq ||v_i - v_i'||_2 9\sqrt{\frac{s}{n}} n^{-1/2} \sqrt{\sum_{i=1}^{s} (u_i^\intercal \mathbf{Y})^2} \\
&\leq ||v_i - v_i'||_2 9\sqrt{\frac{s}{n}} n^{-1/2} ||\mathbf{Y}||_2,
\end{aligned}$$

where the last inequality follows from the fact that the vectors u_i are orthonormal and hence $\sum_{i=1}^{s}(u_i^\intercal \mathbf{Y})^2 \leq ||\mathbf{Y}||_2^2$. Since Y_i are assumed to have finite second moment, by Chebyshev's inequality we have that:

$$\mathbb{P}(|n^{-1}||\mathbf{Y}||_2^2 - \sigma^2| \geq t) \leq \frac{\eta}{nt^2}.$$

Selecting $t = \sqrt{\frac{\log n}{n}}$ is sufficient to keep the above probability going to 0, and furthermore for n large enough guarantees that $n^{-1}||\mathbf{Y}||_2^2 \leq 4\sigma^2$ and hence $n^{-1/2}||\mathbf{Y}||_2 \leq 2\sigma$ (assuming that σ doesn't scale with n). Thus conditional on this event the function F is Lipschitz with a constant equal to $18\sigma\sqrt{\frac{s}{n}}$. Since the expectation of the function F is 0, by concentration of measure for Lipschitz functions on the sphere[43], for any $t > 0$ we have:

$$\mathbb{P}(|F(v_i)| \geq t\sigma) \leq 2\exp\left(-\widetilde{c}s\frac{t^2}{324\frac{s}{n}}\right),$$

for some $\widetilde{c} > 0$. Taking a union bound the above becomes:

$$\mathbb{P}(\max_i |F(v_i)| \geq t\sigma) \leq 2\exp\left(\log(s) - \widetilde{c}\frac{t^2 n}{324}\right).$$

Selecting $t = 26\sqrt{\frac{\log s}{\widetilde{c}n}}$, keeps the probability vanishing at rate faster than $2/s$ and completes the proof. □

C

Proofs for Chapter 4

C.1 Proofs of the General Theory

Proof of Theorem 4.3.3. By the mean value theorem we have:

$$\widehat{S}(\widehat{\boldsymbol{\beta}}_0) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{v}^{*T}\mathbf{h}(\boldsymbol{X}_i, \boldsymbol{\beta}^*) + \underbrace{\frac{1}{n}\sum_{i=1}^{n} \widehat{\mathbf{v}}^T\mathbf{H}(\boldsymbol{X}_i, \widetilde{\boldsymbol{\beta}}_\nu)(\widehat{\boldsymbol{\beta}}_0 - \boldsymbol{\beta}^*)}_{I_1}$$
$$+ \underbrace{\frac{1}{n}\sum_{i=1}^{n}(\widehat{\mathbf{v}} - \mathbf{v}^*)^T\mathbf{h}(\boldsymbol{X}_i, \boldsymbol{\beta}^*)}_{I_2}$$

Next we control I_1:

$$|I_1| \leq \left\| \left[\widehat{\mathbf{v}}^T\mathbf{H}(\boldsymbol{X}_i, \widetilde{\boldsymbol{\beta}}_\nu)\right]_{-1} \right\|_\infty \|\widehat{\boldsymbol{\beta}}_0 - \boldsymbol{\beta}^*\|_1 \leq O_p(r_4(n))O_p(r_1(n)). \tag{C.1.1}$$

where by $[\cdot]_{-1}$ we mean discarding the first entry (corresponding to θ) of the vector. We proceed to bound I_2:

$$|I_2| \leq \|\widehat{\mathbf{v}} - \mathbf{v}^*\|_1 \left\| \frac{1}{n}\sum_{i=1}^{n} \mathbf{h}(\boldsymbol{X}_i, \boldsymbol{\beta}^*) \right\|_\infty = O_p(r_2(n))O_p(r_3(n)). \tag{C.1.2}$$

Thus using (4.3.5) we have:

$$n^{1/2}(|I_1| + |I_2|) \leq n^{1/2}O_p(r_1(n)r_4(n) + r_2(n)r_3(n)) = o_p(1),$$

and we are done. □

Proof of Proposition 4.3.8. Note that the only thing left to show is the consistency of the plugin estimate $\widehat{\mathbf{v}}^T\widehat{\boldsymbol{\Sigma}}\widehat{\mathbf{v}}$ to $\mathbf{v}^{*T}\boldsymbol{\Sigma}\mathbf{v}^*$, with the rest of the argument following from Corollary 4.3.5 and Slutsky's

theorem. By the triangle inequality we have:

$$|\widehat{\mathbf{v}}^T\widehat{\mathbf{\Sigma}}\widehat{\mathbf{v}} - \mathbf{v}^{*T}\mathbf{\Sigma}\mathbf{v}^*| \leq \underbrace{\|\widehat{\mathbf{v}}^T - \mathbf{v}^*\|_1\|\widehat{\mathbf{\Sigma}}(\widehat{\mathbf{v}} - \mathbf{v}^*)\|_\infty}_{I_1} + 2\underbrace{\|\mathbf{v}^{*T}\widehat{\mathbf{\Sigma}}\|_\infty\|\widehat{\mathbf{v}} - \mathbf{v}^*\|_1}_{I_2}$$
$$+ \underbrace{\|\mathbf{v}^*\|_1^2\|\widehat{\mathbf{\Sigma}} - \mathbf{\Sigma}\|_{\max}}_{I_3}$$

Next we control I_1:

$$|I_1| \leq O_p(r_2(n)^2 r_5(n) + \|\mathbf{\Sigma}\|_{\max} r_2^2(n)) = o_p(1).$$

Below we tackle I_2:

$$|I_2| \leq 2O_p(\|\mathbf{v}^*\|_1 r_2(n) r_5(n) + \|\mathbf{v}^{*T}\mathbf{\Sigma}\|_\infty r_2(n)) = o_p(1)$$

Finally, for I_3 we have:

$$|I_3| \leq \|\mathbf{v}^*\|_1^2 O_p(r_5(n)) = o_p(1).$$

□

Proof of Proposition 4.3.26. It is easy to see that, with the help of the mean value theorem, (4.3.21) can be rewritten as:

$$n^{1/2}(\widetilde{\theta} - \theta^*) = n^{1/2}\frac{\overbrace{(\widehat{\theta} - \theta^*)\frac{1}{n}\widehat{\mathbf{v}}^T\sum_{i=1}^n\left(\left[\mathbf{H}(\boldsymbol{X}_i, \widehat{\boldsymbol{\beta}})\right]_{*1} - \left[\mathbf{H}(\boldsymbol{X}_i, \widetilde{\boldsymbol{\beta}}_\nu)\right]_{*1}\right)}^{I_1} - \overbrace{\frac{1}{n}\sum_{i=1}^n \widehat{\mathbf{v}}^T\mathbf{h}(\boldsymbol{X}_i, \widehat{\beta}_{\theta^*})}^{I_3}}{\underbrace{\frac{1}{n}\sum_{i=1}^n \widehat{\mathbf{v}}^T\left[\mathbf{H}(\boldsymbol{X}_i, \widehat{\boldsymbol{\beta}})\right]_{*1}}_{I_2}},$$

where $\widetilde{\boldsymbol{\beta}}_\nu = \nu\widehat{\boldsymbol{\beta}} + (1-\nu)\widehat{\boldsymbol{\beta}}_{\theta^*}$. By Assumption 4.3.25

$$|n^{1/2}(\widehat{\theta} - \theta^*)I_1| \leq n^{1/2}|\widehat{\theta} - \theta^*|O_p(r_5(n)) = o_p(1)$$

Moreover, by Assumption 4.3.25 $|I_2| = 1 + o_p(1)$. By similar arguments to the one we used to show Corollary 4.3.5, we know that $I_3 \rightsquigarrow N(0,1)$. Hence, putting everything together with Slutsky's theorem, we have:

$$\frac{n^{1/2}}{\sqrt{\mathbf{v}^{*T}\boldsymbol{\Sigma}\mathbf{v}^*}}(\widetilde{\theta} - \theta^*) \rightsquigarrow N(0,1),$$

as claimed. □

Lemma C.1.1. *Under Assumptions 4.3.10 — 4.3.13, we have:*

$$\lim_{n\to\infty} \inf_{\boldsymbol{\beta}\in\boldsymbol{\Omega}_0} \mathbb{P}_{\boldsymbol{\beta}}\left(\left|\widehat{S}(0,\widehat{\boldsymbol{\gamma}}) - S(0,\boldsymbol{\gamma})\right| \leq r_1(n)r_4(n) + r_2(n)r_3(n)\right) = 1 \tag{C.1.3}$$

If in addition $n^{1/2}(r_1(n)r_4(n) + r_2(n)r_3(n)) = o(1)$*, we have:*

$$\lim_{n\to\infty} \sup_{\boldsymbol{\beta}\in\boldsymbol{\Omega}_0} \sup_t \left|\mathbb{P}_{\boldsymbol{\beta}}\left(\frac{n^{1/2}}{\sqrt{\mathbf{v}^T\boldsymbol{\Sigma}\mathbf{v}}}\widehat{S}(0,\widehat{\boldsymbol{\gamma}}) \leq t\right) - \Phi(t)\right| = 0 \tag{C.1.4}$$

Proof of Lemma C.1.1. The proof of (C.1.3) is exactly the same as the proof of Theorem 4.3.3, but uses the uniform convergence assumptions. Note that the bounds, (C.1.1) and (C.1.2) still hold as long as the event $\mathcal{G}^{\boldsymbol{\beta}} = \mathcal{G}_1^{\boldsymbol{\beta}} \cap \ldots \cap \mathcal{G}_4^{\boldsymbol{\beta}}$ holds. Since $\inf_{\boldsymbol{\beta}\in\boldsymbol{\Omega}_0} \mathbb{P}_{\boldsymbol{\beta}}(\mathcal{G}^{\boldsymbol{\beta}}) \geq 1 - \sum_{i=1}^4 \sup_{\boldsymbol{\beta}\in\boldsymbol{\Omega}_0} \mathbb{P}_{\boldsymbol{\beta}}[(\mathcal{G}_i^{\boldsymbol{\beta}})^c] \to 1$ by Assumptions 4.3.10 — 4.3.13, this completes the proof of (C.1.3).

Next we show (C.1.4). Let $\kappa(n) = \sqrt{n}C^{-1/2}(r_1(n)r_4(n) + r_2(n)r_3(n))$, where we recall the

definition of C: $C = \inf_{\boldsymbol{\beta}\in\boldsymbol{\Omega}_0} \mathbf{v}^T\boldsymbol{\Sigma}\mathbf{v} > 0$. Then we have:

$$\begin{aligned}\mathbb{P}_{\boldsymbol{\beta}}\left(\frac{n^{1/2}}{\sqrt{\mathbf{v}^T\boldsymbol{\Sigma}\mathbf{v}}}\widehat{S}(0,\widehat{\boldsymbol{\gamma}}) \le t\right) &\le \mathbb{P}_{\boldsymbol{\beta}}\left(\frac{n^{1/2}}{\sqrt{\mathbf{v}^T\boldsymbol{\Sigma}\mathbf{v}}}\widehat{S}(0,\widehat{\boldsymbol{\gamma}}) \le t, \mathcal{G}^{\boldsymbol{\beta}}\right) + \mathbb{P}_{\boldsymbol{\beta}}((\mathcal{G}^{\boldsymbol{\beta}})^c) \\ &\le \mathbb{P}_{\boldsymbol{\beta}}\left(\frac{n^{1/2}}{\sqrt{\mathbf{v}^T\boldsymbol{\Sigma}\mathbf{v}}}S(0,\boldsymbol{\gamma}) \le t + \kappa(n)\right) + \mathbb{P}_{\boldsymbol{\beta}}((\mathcal{G}^{\boldsymbol{\beta}})^c)\end{aligned}$$

The above implies the following inequality:

$$\begin{aligned}&\mathbb{P}_{\boldsymbol{\beta}}\left(\frac{n^{1/2}}{\sqrt{\mathbf{v}^T\boldsymbol{\Sigma}\mathbf{v}}}\widehat{S}(0,\widehat{\boldsymbol{\gamma}}) \le t\right) - \Phi(t) \\ &\le \mathbb{P}_{\boldsymbol{\beta}}\left(\frac{n^{1/2}}{\sqrt{\mathbf{v}^T\boldsymbol{\Sigma}\mathbf{v}}}S(0,\boldsymbol{\gamma}) \le t + \kappa(n)\right) - \Phi(t+\kappa(n)) + (\Phi(t+\kappa(n)) - \Phi(t)) + \mathbb{P}_{\boldsymbol{\beta}}((\mathcal{G}^{\boldsymbol{\beta}})^c) \\ &\le \mathbb{P}_{\boldsymbol{\beta}}\left(\frac{n^{1/2}}{\sqrt{\mathbf{v}^T\boldsymbol{\Sigma}\mathbf{v}}}S(0,\boldsymbol{\gamma}) \le t + \kappa(n)\right) - \Phi(t+\kappa(n)) + \frac{\kappa(n)}{\sqrt{2\pi}} + \mathbb{P}_{\boldsymbol{\beta}}((\mathcal{G}^{\boldsymbol{\beta}})^c),\end{aligned}$$

where we took into account the fact that Φ is Lipschitz with constant $\le \frac{1}{\sqrt{2\pi}}$. Now taking into account Assumption 4.3.12, the fact that $\kappa(n) = o(1)$ and $\mathbb{P}_{\boldsymbol{\beta}}((\mathcal{G}^{\boldsymbol{\beta}})^c) = o(1)$ we conclude that:

$$\limsup_{n\to\infty} \sup_{\boldsymbol{\beta}\in\boldsymbol{\Omega}_0} \sup_t \mathbb{P}_{\boldsymbol{\beta}}\left(\frac{n^{1/2}}{\sqrt{\mathbf{v}^T\boldsymbol{\Sigma}\mathbf{v}}}\widehat{S}(0,\widehat{\boldsymbol{\gamma}}) \le t\right) - \Phi(t) \le 0.$$

With a similar argument one can show the reverse, namely:

$$\liminf_{n\to\infty} \inf_{\boldsymbol{\beta}\in\boldsymbol{\Omega}_0} \inf_t \mathbb{P}_{\boldsymbol{\beta}}\left(\frac{n^{1/2}}{\sqrt{\mathbf{v}^T\boldsymbol{\Sigma}\mathbf{v}}}\widehat{S}(0,\widehat{\boldsymbol{\gamma}}) \le t\right) - \Phi(t) \ge 0.$$

This concludes the proof of (C.1.4). □

Proof of Theorem 4.3.14. Let $\mathcal{G}^{\boldsymbol{\beta}} = \mathcal{G}_1^{\boldsymbol{\beta}} \cap \ldots \cap \mathcal{G}_5^{\boldsymbol{\beta}}$. By Assumption 4.3.13, on the event $\mathcal{G}^{\boldsymbol{\beta}}$ we

clearly have:

$$|\widehat{\sigma}^2 - \mathbf{v}^T \boldsymbol{\Sigma} \mathbf{v}| \leq \tau(n) \tag{C.1.5}$$

Note that (C.1.5) immediately implies that $\widehat{\sigma} \geq \sqrt{\mathbf{v}^T \boldsymbol{\Sigma} \mathbf{v} - \tau(n)} \geq \sqrt{C - \tau(n)}$, by assumption. Hence note that

$$|\widehat{\sigma} - \sqrt{\mathbf{v}^T \boldsymbol{\Sigma} \mathbf{v}}| \leq \frac{\tau(n)}{\sqrt{C - \tau(n)} + \sqrt{C}} \leq \frac{\tau(n)}{\sqrt{C - \tau(n)}} \tag{C.1.6}$$

We next investigate the following difference, for some $\alpha > 0$:

$$\begin{aligned}
&\underbrace{\mathbb{P}_{\boldsymbol{\beta}}\left(\frac{n^{1/2}}{\widehat{\sigma}} \widehat{S}(0, \widehat{\boldsymbol{\gamma}}) \leq t\right) - \mathbb{P}_{\boldsymbol{\beta}}\left(\frac{n^{1/2}}{\sqrt{\mathbf{v}^T \boldsymbol{\Sigma} \mathbf{v}}} \widehat{S}(0, \widehat{\boldsymbol{\gamma}}) \leq t + \alpha\right)}_{I_1} \\
&\leq \mathbb{P}_{\boldsymbol{\beta}}\left(\left|\frac{n^{1/2}}{\sqrt{\mathbf{v}^T \boldsymbol{\Sigma} \mathbf{v}}} \widehat{S}(0, \widehat{\boldsymbol{\gamma}})\right| \frac{|\widehat{\sigma} - \sqrt{\mathbf{v}^T \boldsymbol{\Sigma} \mathbf{v}}|}{\widehat{\sigma}} > \alpha\right) \\
&\leq \mathbb{P}_{\boldsymbol{\beta}}\left(\left|\frac{n^{1/2}}{\sqrt{\mathbf{v}^T \boldsymbol{\Sigma} \mathbf{v}}} \widehat{S}(0, \widehat{\boldsymbol{\gamma}})\right| > \alpha^{-1}\right) + \mathbb{P}_{\boldsymbol{\beta}}\left(\frac{|\widehat{\sigma} - \sqrt{\mathbf{v}^T \boldsymbol{\Sigma} \mathbf{v}}|}{\widehat{\sigma}} > \alpha^2, \mathcal{G}^{\boldsymbol{\beta}}\right) + \mathbb{P}_{\boldsymbol{\beta}}((\mathcal{G}^{\boldsymbol{\beta}})^c)
\end{aligned}$$

From Lemma C.1.1, we know that for any α (even depending on n or $\boldsymbol{\beta}$):

$$\lim_{n \to \infty} \sup_{\boldsymbol{\beta} \in \boldsymbol{\Omega}_0} \sup_{\alpha} \left| \mathbb{P}_{\boldsymbol{\beta}}\left(\left|\frac{n^{1/2}}{\sqrt{\mathbf{v}^T \boldsymbol{\Sigma} \mathbf{v}}} \widehat{S}(0, \widehat{\boldsymbol{\gamma}})\right| > \alpha^{-1}\right) - \mathbb{P}(|Z| \geq \alpha^{-1}))\right| = 0,$$

where $Z \sim N(0, 1)$, and hence as $n \to \infty$:

$$\mathbb{P}_{\boldsymbol{\beta}}\left(\left|\frac{n^{1/2}}{\sqrt{\mathbf{v}^T \boldsymbol{\Sigma} \mathbf{v}}} \widehat{S}(0, \widehat{\boldsymbol{\gamma}})\right| > \alpha^{-1}\right) \leq 2(1 - \Phi(\alpha^{-1})) \leq \frac{2}{\sqrt{2\pi}} \alpha \exp\left(-\frac{1}{2\alpha^2}\right) \to 0$$

by a standard tail bound for the standard normal cdf. Next we deal with:

$$\mathbb{P}_{\boldsymbol{\beta}}\left(\frac{|\widehat{\sigma}-\sqrt{\mathbf{v}^T\boldsymbol{\Sigma}\mathbf{v}}|}{\widehat{\sigma}}>\alpha^2,\mathcal{G}^{\boldsymbol{\beta}}\right)\leq\mathbb{P}_{\boldsymbol{\beta}}\left(\frac{\tau(n)}{C-\tau(n)}>\alpha^2\right)\leq\mathbb{P}_{\boldsymbol{\beta}}\left(\frac{2\tau(n)}{C}>\alpha^2\right),$$

where we used the bound (C.1.6), and we are assuming n is large enough so $\tau(n)<C/2$ e.g. Taking $\alpha=\sqrt[3]{\tau(n)}$ is sufficient to let the above probability converge to 0, while keeping $\alpha\to 0$. Note further that due to Assumptions 4.3.10 — 4.3.13, we have $\lim_{n\to\infty}\sup_{\beta\in\boldsymbol{\Omega}_0}\mathbb{P}_{\boldsymbol{\beta}}((\mathcal{G}^{\boldsymbol{\beta}})^c)=0$. Hence we have shown $\limsup_{n\to\infty}\sup_{\boldsymbol{\beta}\in\boldsymbol{\Omega}_0}\sup_t I_1\leq 0$. Finally note that the following decomposition evidently holds:

$$\underbrace{\mathbb{P}_{\boldsymbol{\beta}}\left(\frac{n^{1/2}}{\widehat{\sigma}}\widehat{S}(0,\widehat{\boldsymbol{\gamma}})\leq t\right)-\Phi(t)}_{I}=I_1+\underbrace{\mathbb{P}_{\boldsymbol{\beta}}\left(\frac{n^{1/2}}{\sqrt{\mathbf{v}^T\boldsymbol{\Sigma}\mathbf{v}}}\widehat{S}(0,\widehat{\boldsymbol{\gamma}})\leq t+\alpha\right)-\Phi(t+\alpha)}_{I_2}+\underbrace{\Phi(t)-\Phi(t+\alpha)}_{I_3}.$$

Using arguments as in Lemma C.1.1, we can show that $\limsup_{n\to\infty}\sup_{\boldsymbol{\beta}\in\boldsymbol{\Omega}_0}\sup_t\max(I_2,I_3)\leq 0$, since as we mentioned $\alpha\to 0$. Hence:

$$\limsup_{n\to\infty}\sup_{\boldsymbol{\beta}\in\boldsymbol{\Omega}_0}\sup_t I\leq 0.$$

Analogously we can show that:

$$\liminf_{n\to\infty}\inf_{\boldsymbol{\beta}\in\boldsymbol{\Omega}_0}\inf_t I\geq 0,$$

which completes the proof. □

Lemma C.1.2. *Under Assumptions 4.3.17 — 4.3.19 and Assumption 4.3.21 we have that:*

$$\lim_{n\to\infty}\sup_{\boldsymbol{\beta}\in\boldsymbol{\Omega}_1(K,\phi)}\mathbb{P}_{\boldsymbol{\beta}}(\sqrt{n}|\widehat{S}(0,\widehat{\boldsymbol{\gamma}})-S(\boldsymbol{\beta})+\theta|\leq r_6(n)+\sqrt{n}\kappa(n))=1, \tag{C.1.7}$$

where $\kappa(n) = r_1(n)r_4(n) + r_2(n)r_3(n)$. *Furthermore, assuming that* $\sqrt{n}\kappa(n) = o(1)$, *we have:*

$$\lim_{n\to\infty} \sup_{\boldsymbol{\beta}\in\boldsymbol{\Omega}_1(K,\phi)} \sup_t \left| \mathbb{P}_{\boldsymbol{\beta}} \left(\frac{n^{1/2}}{\sqrt{\mathbf{v}^T\boldsymbol{\Sigma}\mathbf{v}}} \widehat{S}(0,\widehat{\boldsymbol{\gamma}}) \leq t \right) - \Phi(t) \right| = 0, \ \textit{if}\, \phi > 1/2, \tag{C.1.8}$$

$$\lim_{n\to\infty} \sup_{\boldsymbol{\beta}\in\boldsymbol{\Omega}_1(K,\phi)} \sup_t \left| \mathbb{P}_{\boldsymbol{\beta}} \left(\frac{n^{1/2}}{\sqrt{\mathbf{v}^T\boldsymbol{\Sigma}\mathbf{v}}} \widehat{S}(0,\widehat{\boldsymbol{\gamma}}) \leq t \right) - \Phi\left(t + \frac{K}{\sqrt{\mathbf{v}^T\boldsymbol{\Sigma}\mathbf{v}}} \right) \right| = 0, \ \textit{if}\, \phi = 1/2, \tag{C.1.9}$$

and for a fixed $t \in \mathbb{R}$ *and* $K \neq 0$ *we have:*

$$\lim_{n\to\infty} \sup_{\boldsymbol{\beta}\in\boldsymbol{\Omega}_1(K,\phi)} \mathbb{P}_{\boldsymbol{\beta}} \left(\left| \frac{n^{1/2}}{\sqrt{\mathbf{v}^T\boldsymbol{\Sigma}\mathbf{v}}} \widehat{S}(0,\widehat{\boldsymbol{\gamma}}) \right| \leq t \right) = 0, \ \textit{if}\, \phi < 1/2. \tag{C.1.10}$$

Proof of Lemma C.1.2. To show (C.1.7) we note that by the triangle inequality:

$$|\widehat{S}(0,\widehat{\boldsymbol{\gamma}}) - S(\boldsymbol{\beta}) + \theta| \leq |\widehat{S}(0,\widehat{\boldsymbol{\gamma}}) - S(0,\boldsymbol{\gamma})| + |S(0,\boldsymbol{\gamma}) - S(\boldsymbol{\beta}) + \theta|$$

On the event $\mathcal{G}^{\boldsymbol{\beta}} = \mathcal{G}_1^{\boldsymbol{\beta}} \cap \mathcal{G}_2^{\boldsymbol{\beta}} \cap \mathcal{G}_3^{\boldsymbol{\beta}} \cap \mathcal{G}_4^{\boldsymbol{\beta}} \cap \mathcal{G}_6^{\boldsymbol{\beta}}$ the bounds (C.1.1) and (C.1.2) derived in Theorem 4.3.3 hold. Since $\inf_{\boldsymbol{\beta}\in\boldsymbol{\Omega}_1(K,\phi)} \mathbb{P}_{\boldsymbol{\beta}}(\mathcal{G}^{\boldsymbol{\beta}}) \geq 1 - \sum_{i=1,i\neq 5}^{6} \sup_{\boldsymbol{\beta}\in\boldsymbol{\Omega}_1(K,\phi)} \mathbb{P}_{\boldsymbol{\beta}}[(\mathcal{G}_i^{\boldsymbol{\beta}})^c] \to 1$, we have that $|\widehat{S}(0,\widehat{\boldsymbol{\gamma}}) - S(0,\boldsymbol{\gamma})| \leq \kappa(n)$, using Assumption 4.3.21 completes the proof of (C.1.7).

Let $\xi(n) = C^{-1/2}(\sqrt{n}\kappa(n) + r_6(n))$, where we recall the definition of C: $C = \inf_{\boldsymbol{\beta}\in\boldsymbol{\Omega}_1(K,\phi)} \mathbf{v}^T\boldsymbol{\Sigma}\mathbf{v} > 0$. Furthermore denote for brevity $\mu := \frac{1}{\sqrt{\mathbf{v}^T\boldsymbol{\Sigma}\mathbf{v}}}$. Using (C.1.7) we have:

$$\begin{aligned}
&\mathbb{P}_{\boldsymbol{\beta}} \left(\frac{n^{1/2}}{\sqrt{\mathbf{v}^T\boldsymbol{\Sigma}\mathbf{v}}} \widehat{S}(0,\widehat{\boldsymbol{\gamma}}) \leq t \right) - \Phi(t) \\
&\leq \mathbb{P}_{\boldsymbol{\beta}} \left(\frac{n^{1/2}}{\sqrt{\mathbf{v}^T\boldsymbol{\Sigma}\mathbf{v}}} S(\boldsymbol{\beta}) \leq t + \xi(n) + Kn^{1/2-\phi}\mu \right) - \Phi(t + \xi(n) + Kn^{1/2-\phi}\mu) + \mathbb{P}_{\boldsymbol{\beta}}((\mathcal{G}^{\boldsymbol{\beta}})^c) \\
&\quad + \Phi(t + \xi(n) + Kn^{1/2-\phi}\mu) - \Phi(t)
\end{aligned} \tag{C.1.11}$$

Recall that $\Phi(t+\xi(n)+Kn^{1/2-\phi}\mu)-\Phi(t) \leq \frac{\xi(n)+Kn^{1/2-\phi}\mu}{\sqrt{2\pi}}$, and by Assumption 4.3.19, (C.1.11) gives:

$$\limsup_{n\to\infty} \sup_{\boldsymbol{\beta}\in\boldsymbol{\Omega}_1(K,\phi)} \sup_t \mathbb{P}_{\boldsymbol{\beta}}\left(\frac{n^{1/2}}{\sqrt{\mathbf{v}^T\boldsymbol{\Sigma}\mathbf{v}}}\widehat{S}(0,\widehat{\boldsymbol{\gamma}}) \leq t\right) - \Phi(t) \leq 0$$

Using similar arguments we can obtain the lower bound and conclude the proof of (C.1.8). The proof of (C.1.9) follows analogously.

Finally we show (C.1.10). Denote with $l := -t - \xi(n) + Kn^{1/2-\phi}\mu$ and $L := t + \xi(n) + Kn^{1/2-\phi}\mu$, and similarly to the bound in (C.1.11) we have:

$$\begin{aligned}
&\mathbb{P}_{\boldsymbol{\beta}}\left(\left|\frac{n^{1/2}}{\sqrt{\mathbf{v}^T\boldsymbol{\Sigma}\mathbf{v}}}\widehat{S}(0,\widehat{\boldsymbol{\gamma}})\right| \leq t\right) \\
&\leq \mathbb{P}_{\boldsymbol{\beta}}\left(l \leq \frac{n^{1/2}}{\sqrt{\mathbf{v}^T\boldsymbol{\Sigma}\mathbf{v}}}S(\boldsymbol{\beta}) \leq L\right) - \mathbb{P}(l \leq Z \leq L) + \mathbb{P}(l \leq Z \leq L) + \mathbb{P}_{\boldsymbol{\beta}}((\mathcal{G}^{\boldsymbol{\beta}})), \quad \text{(C.1.12)}
\end{aligned}$$

where $Z \sim N(0,1)$. Note that since $\phi < 1/2, l, L \to \infty$, if $K > 0$ $\mathbb{P}(l \leq Z \leq L) = \Phi(L) - \Phi(l) \to 0$. Similarly if $K < 0$ we have $\mathbb{P}(l \leq Z \leq L) \leq \Phi(L) \to 0$. Moreover by Assumption 4.3.19 we know that:

$$\limsup_{n\to\infty} \sup_{\boldsymbol{\beta}\in\boldsymbol{\Omega}_1(K,\phi)} \left|\mathbb{P}_{\boldsymbol{\beta}}\left(l \leq \frac{n^{1/2}}{\sqrt{\mathbf{v}^T\boldsymbol{\Sigma}\mathbf{v}}}S(\boldsymbol{\beta}) \leq L\right) - \mathbb{P}(l \leq Z \leq L)\right| = 0,$$

which completes the proof. □

Proof of Theorem 4.3.22. Since the proofs of (4.3.17) and (4.3.18) are very similar we will only show (4.3.18). Note that on the event $\mathcal{G}^{\boldsymbol{\beta}} = \mathcal{G}_1^{\boldsymbol{\beta}} \cap \ldots \cap \mathcal{G}_6^{\boldsymbol{\beta}}$, just as in the proof Theorem 4.3.14, we have that the bounds (C.1.5) and (C.1.6). We have that $\sup_{\boldsymbol{\beta}\in\boldsymbol{\Omega}_1(K,\phi)} \tau(n) = o(1)$. Set $U_n := \frac{n^{1/2}}{\sqrt{\mathbf{v}^T\boldsymbol{\Sigma}\mathbf{v}}}\widehat{S}(0,\widehat{\boldsymbol{\gamma}})$ and $\mu := \frac{1}{\sqrt{\mathbf{v}^T\boldsymbol{\Sigma}\mathbf{v}}}$. The proof then proceeds similarly to the proof of Theorem

4.3.14. Decompose:

$$\mathbb{P}_{\boldsymbol{\beta}}\left(\widehat{U}_n \leq t\right) - \Phi\left(t + K\mu\right) = \underbrace{\mathbb{P}_{\boldsymbol{\beta}}\left(\widehat{U}_n \leq t\right) - \mathbb{P}_{\boldsymbol{\beta}}\left(U_n \leq t + \alpha\right)}_{I_1}$$
$$+ \underbrace{\mathbb{P}_{\boldsymbol{\beta}}\left(U_n \leq t + \alpha\right) - \Phi(t + K\mu + \alpha)}_{I_2} + \underbrace{\Phi(t + K\mu) - \Phi(t + K\mu + \alpha)}_{I_3}.$$

Starting from I_1, for some $\alpha > 0$ we have :

$$\sup_{t\in\mathbb{R}} I_1 \leq \mathbb{P}_{\boldsymbol{\beta}}(|\widehat{U}_n - U_n| \geq \alpha) = \mathbb{P}_{\boldsymbol{\beta}}\left(|U_n|\left|1 - \frac{\sqrt{\mathbf{v}^T\boldsymbol{\Sigma}\mathbf{v}}}{\widehat{\sigma}}\right| \geq \alpha\right)$$
$$\leq \mathbb{P}_{\boldsymbol{\beta}}(|U_n| \geq \alpha^{-1}) + \mathbb{P}_{\boldsymbol{\beta}}\left(\left|1 - \frac{\sqrt{\mathbf{v}^T\boldsymbol{\Sigma}\mathbf{v}}}{\widehat{\sigma}}\right| \geq \alpha^2\right) \quad \text{(C.1.13)}$$

For the first term in (C.1.13), we have:

$$\mathbb{P}_{\boldsymbol{\beta}}(|U_n| \geq \alpha^{-1}) \leq |\mathbb{P}_{\boldsymbol{\beta}}(|U_n| \geq \alpha^{-1}) - \mathbb{P}(|Z - K\mu| \geq \alpha^{-1})| + \mathbb{P}(|Z - K\mu| \geq \alpha^{-1}),$$

where $Z \sim N(0,1)$. By Lemma C.1.2:

$$\lim_{n\to\infty} \sup_{\boldsymbol{\beta}\in\boldsymbol{\Omega}_1(K,\phi)} \sup_{\alpha} |\mathbb{P}_{\boldsymbol{\beta}}(|U_n| \geq \alpha^{-1}) - \mathbb{P}(|Z - K\mu| \geq \alpha^{-1})| = 0$$

Furthermore, by a standard tail bound:

$$\mathbb{P}(|Z - K\mu| \geq \alpha^{-1}) \leq \frac{2}{\sqrt{2\pi}(\alpha^{-1} + K\mu)} \exp\left(-\frac{(\alpha^{-1} + K\mu)^2}{2}\right) \to 0,$$

as $\alpha \to 0$, where the convergence holds uniformly in $\boldsymbol{\beta}$, because $\mu \leq C^{-1}$, by Assumption 4.3.19.

Showing that the second term in (C.1.13) converges to 0, can be done by choosing $\alpha = \sqrt[3]{\tau(n)}$

and using the same technique as the end of Theorem 4.3.14, so we omit it. This implies that:

$$\limsup_{n\to\infty} \sup_{\boldsymbol{\beta}\in\boldsymbol{\Omega}_1(K,\phi)} \sup_t I_1 \leq 0.$$

By Lemma C.1.2, we have:

$$\limsup_{n\to\infty} \sup_{\boldsymbol{\beta}\in\boldsymbol{\Omega}_1(K,\phi)} \sup_t I_2 \leq 0,$$

and since $I_3 \leq (2\pi)^{-1/2}\alpha$ we get:

$$\limsup_{n\to\infty} \sup_{\boldsymbol{\beta}\in\boldsymbol{\Omega}_1(K,\phi)} \sup_t \mathbb{P}_{\boldsymbol{\beta}}\left(\widehat{U}_n \leq t\right) - \Phi\left(t + K\mu\right) \leq 0$$

Similarly we can obtain the lower bound, and conclude the proof of (4.3.18).

For showing (4.3.19), note that since $\sup_{\boldsymbol{\beta}\in\boldsymbol{\Omega}_1(K,\phi)} |\widehat{\sigma}^2 - \mathbf{v}^T\boldsymbol{\Sigma}\mathbf{v}| = o_p(1)$ and by assumption $\inf_{\boldsymbol{\beta}\in\boldsymbol{\Omega}_1(K,\phi)} \mathbf{v}^T\boldsymbol{\Sigma}\mathbf{v}^T \geq C$, we have: $\sup_{\boldsymbol{\beta}\in\boldsymbol{\Omega}_1(K,\phi)} |\widehat{\sigma}^2/\mathbf{v}^T\boldsymbol{\Sigma}\mathbf{v} - 1| \leq 1$, for large n. Thus for a fixed t, for large enough n, we have:

$$\mathbb{P}_{\boldsymbol{\beta}}(|\widehat{U}_n| \leq t) = \mathbb{P}_{\boldsymbol{\beta}}(|U_n| \leq t(\widehat{\sigma}^2/\mathbf{v}^T\boldsymbol{\Sigma}\mathbf{v})^{1/2}) \leq \mathbb{P}_{\beta}(|U_n| \leq \sqrt{2}t).$$

Finally, applying Lemma C.1.2, we get:

$$\lim_{n\to\infty} \sup_{\boldsymbol{\beta}\in\boldsymbol{\Omega}_1(K,\theta)} \mathbb{P}_{\boldsymbol{\beta}}(|\widehat{U}_n| \leq t) \leq \lim_{n\to\infty} \sup_{\boldsymbol{\beta}\in\boldsymbol{\Omega}_1(K,\theta)} \mathbb{P}_{\beta}(|U_n| \leq \sqrt{2}t) = 0,$$

which concludes the proof. □

C.2 Proofs for the Dantzig Selector

Proof of Theorem 4.4.1. Assume that $\boldsymbol{X}_i = (X_{i1}, \boldsymbol{X}_{i,-1}^T)^T$ corresponding to the partition of $\boldsymbol{\beta} = (\theta, \boldsymbol{\gamma})$. The test statistic $\widehat{S}(0, \widehat{\boldsymbol{\gamma}})$ is

$$\widehat{S}(0, \widehat{\boldsymbol{\gamma}}) \;=\; \frac{1}{n} \sum_{i=1}^{n} \widehat{\mathbf{v}}^T \boldsymbol{X}_i (\boldsymbol{X}_{i,-1}^T \widehat{\boldsymbol{\gamma}} - Y_i).$$

Next we verify the assumptions of Theorem 4.3.3. From Lemma C.2.5, we have that $\|\mathbf{v}^* - \widehat{\mathbf{v}}\|_1 = O_p\left(\|\mathbf{v}^*\|_1 s_{\mathbf{v}} \sqrt{\frac{\log d}{n}}\right)$, and within the proof of Lemma C.2.6 (see (C.2.9)), we can see that $\left\| \frac{1}{n} \sum_{i=1}^n \boldsymbol{X}_i (\boldsymbol{X}_{i,-1}^T \boldsymbol{\gamma}^* - Y_i) \right\|_\infty = O_p\left(\sqrt{\frac{\log d}{n}}\right)$, which implies that:

$$n^{1/2} O_p\left(\|\mathbf{v}^*\|_1 s_{\mathbf{v}} \sqrt{\frac{\log d}{n}}\right) O_p\left(\sqrt{\frac{\log d}{n}}\right) = O_p\left(\|\mathbf{v}^*\|_1 s_{\mathbf{v}} \frac{\log d}{\sqrt{n}}\right) = o_p(1). \tag{C.2.1}$$

Furthermore, by Lemma C.2.7, we have that $\|\boldsymbol{\beta}^* - \widehat{\boldsymbol{\beta}}\|_1 = O_p\left(s\sqrt{\frac{\log d}{n}}\right)$, and since we know that it suffices to select $\lambda' = \widetilde{C}\|\mathbf{v}^*\|_1 \sqrt{\frac{\log d}{n}}$, for some large constant $\widetilde{C}$ (see Lemma C.2.5), we get

$$\sqrt{n}\lambda' \|\boldsymbol{\beta}^* - \widehat{\boldsymbol{\beta}}\|_1 = \sqrt{n}\widetilde{C}\|\mathbf{v}^*\|_1 \sqrt{\frac{\log d}{n}} O_p\left(s\sqrt{\frac{\log d}{n}}\right) = o_p(1). \tag{C.2.2}$$

Adding up (C.2.1) and (C.2.2) yields condition (4.3.5) from Theorem 4.3.3. Hence, the influence function expansion of $n^{1/2}\widehat{S}(0, \widehat{\boldsymbol{\gamma}})$ follows from Theorem 4.3.3. □

Proof of Corollary 4.4.5. The influence function expansion from Theorem 4.4.1, has already given us a decomposition of $n^{1/2}\widehat{S}(0, \widehat{\boldsymbol{\gamma}})$ into iid terms with mean 0. We next show that the conditions from Corollary 4.3.5 hold by verifying Lyapunov's condition for the CLT — which will verify As-

sumption 4.3.4. We need to show that the following expression converges to 0:

$$\frac{n^{-3/2}}{\Delta^{3/2}} \sum_{i=1}^{n} \mathbb{E} \left| \mathbf{v}^{*T} \boldsymbol{X}_i (\boldsymbol{X}_{i,-1}^T \boldsymbol{\gamma}^* - Y_i) \right|^3 .$$

Note that we have $\Delta^{3/2} \geq \lambda_{\min}(\boldsymbol{\Sigma}_{\boldsymbol{X}}) \|\mathbf{v}^*\|_2^3 \operatorname{Var}(\varepsilon)^{3/2} = O(1)\|\mathbf{v}^*\|_2^3$. Therefore it suffices to consider the following expression:

$$\begin{aligned} \frac{n^{-3/2}}{\|\mathbf{v}^*\|_2^3} \sum_{i=1}^{n} \mathbb{E} \left| \mathbf{v}^{*T} \boldsymbol{X}_i (\boldsymbol{X}_{i,-1}^T \boldsymbol{\gamma}^* - Y_i) \right|^3 &\leq n^{-3/2} \sum_{i=1}^{n} \mathbb{E} \, \|(\boldsymbol{X}_i \varepsilon_i)_{S_{\mathbf{v}}}\|_2^3 \\ &\leq n^{-1/2} s_{\mathbf{v}}^{3/2} M, \end{aligned} \tag{C.2.3}$$

where $M = (6KK_{\boldsymbol{X}})^3$, and the last inequality holding from Lemma C.2.9. This completes the proof. □

Remark C.2.1. *Using the Berry-Esseen theorem for non-identical random variables in combination with (C.2.3) we can further show:*

$$\sup_t \left| \mathbb{P}^* \left(\frac{n^{1/2}}{\sqrt{\Delta}} S(0, \boldsymbol{\gamma}^*) \leq t \right) - \Phi(t) \right| \leq C_{BE} M n^{-1/2} s_{\mathbf{v}}^{3/2} = o(1),$$

where C_{BE} is an absolute constant.

Proof of Proposition 4.4.6. We show that each of the two sums is corresponding to it's population counterpart, and then the proof follows upon an application of Slutsky's theorem. We start with the first term:

$$\left| \frac{1}{n} \sum_{i=1}^{n} (\widehat{\mathbf{v}}^T \boldsymbol{X}_i)^2 - \mathbf{v}^{*T} \boldsymbol{\Sigma}_{\boldsymbol{X}} \mathbf{v}^* \right| \leq \underbrace{\left| \frac{1}{n} \sum_{i=1}^{n} [(\widehat{\mathbf{v}}^T \boldsymbol{X}_i)^2 - (\mathbf{v}^{*T} \boldsymbol{X}_i)^2] \right|}_{I_1} + \underbrace{|\mathbf{v}^{*T} \boldsymbol{\Sigma}_n \mathbf{v}^* - \mathbf{v}^{*T} \boldsymbol{\Sigma}_{\boldsymbol{X}} \mathbf{v}^*|}_{I_2},$$

$$|I_1| \leq \|\widehat{\mathbf{v}} - \mathbf{v}^*\|_1 (\|\boldsymbol{\Sigma}_n \widehat{\mathbf{v}}\|_\infty + \|\boldsymbol{\Sigma}_n \mathbf{v}^*\|_\infty).$$

We know from Lemma C.2.5, that $\|\mathbf{v}^* - \widehat{\mathbf{v}}\|_1 = O_p\left(\|\mathbf{v}^*\|_1 s_{\mathbf{v}} \sqrt{\frac{\log d}{n}}\right)$, and by definition $\|\boldsymbol{\Sigma}_n \widehat{\mathbf{v}}\|_\infty \leq 1 + \lambda'$. In the proof of Lemma C.2.5 we also show that, $\|\boldsymbol{\Sigma}_n \mathbf{v}^*\|_\infty = 1 + O_p\left(\|\mathbf{v}^*\|_1 \sqrt{\frac{\log d}{n}}\right)$ upon appropriately choosing $\lambda' \asymp \|\mathbf{v}^*\|_1 \sqrt{\frac{\log d}{n}}$, with a large enough proportionality constant. Thus since $O_p\left(\|\mathbf{v}^*\|_1 s_{\mathbf{v}} \sqrt{\frac{\log d}{n}}\right)\left(2 + O_p\left(\|\mathbf{v}^*\|_1 \sqrt{\frac{\log d}{n}}\right)\right) = o_p(1)$ we have shown $|I_1| = o_p(1)$. We next tackle I_2:

$$|I_2| \leq \|\mathbf{v}^*\|_1^2 \|\boldsymbol{\Sigma}_n - \boldsymbol{\Sigma}_{\boldsymbol{X}}\|_{\max}.$$

Lemma C.2.2 gives us that $\|\boldsymbol{\Sigma}_n - \boldsymbol{\Sigma}_{\boldsymbol{X}}\|_{\max} = O_p(\sqrt{\frac{\log d}{n}})$, and thus because of our extra assumption we have $|I_2| = O_p(\|\mathbf{v}^*\|_1^2 \sqrt{\frac{\log d}{n}}) = o_p(1)$.

Now we turn to the second part of the proof:

$$\left|\frac{1}{n}\sum_{i=1}^n (Y_i - \boldsymbol{X}_i^T \widehat{\boldsymbol{\beta}})^2 - \mathrm{Var}(\varepsilon)\right| \leq \underbrace{\left|\frac{1}{n}\sum_{i=1}^n (Y_i - \boldsymbol{X}_i^T \widehat{\boldsymbol{\beta}})^2 - \frac{1}{n}\sum_{i=1}^n (Y_i - \boldsymbol{X}_i^T \boldsymbol{\beta}^*)^2\right|}_{I_3} + \underbrace{\left|\frac{1}{n}\sum_{i=1}^n \varepsilon_i^2 - \mathrm{Var}(\varepsilon)\right|}_{I_4}.$$

The term I_4 is clearly $o_p(1)$ because of the LLN (ε_i are centered and have finite variance as sub-Gaussian random variables). Thus we are left to deal with I_3:

$$\begin{aligned} |I_3| &\leq \frac{1}{n}\|\mathbf{X}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)\|_2^2 + \frac{2}{n}\sum_{i=1}^n |\boldsymbol{X}_i^T(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)||\varepsilon_i| \\ &\leq \frac{1}{n}\|\mathbf{X}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)\|_2^2 + \frac{2}{n}\|\mathbf{X}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)\|_2 \sqrt{\sum_{i=1}^n \varepsilon_i^2}, \end{aligned}$$

where $\mathbf{X}_{n\times d}$ is a matrix, with rows $\boldsymbol{X}_i^T$ stacked together. (C.2.11) in Lemma C.2.7 gives us that $\frac{1}{n}\|\mathbf{X}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)\|_2^2 = O_p(\frac{s \log d}{n}) = o_p(1)$, and since by LLN $\sqrt{\frac{1}{n}\sum_{i=1}^n \varepsilon_i^2} = O_p(1)$, we have

$|I_3| = o_p(1)$, which shows the consistency of the second estimator and concludes the proof. □

Proof of Remark 4.4.7. In the second part of the proof of Proposition 4.4.6 we showed that $n^{-1}\sum_{i=1}^n (Y_i - \boldsymbol{X}_i^T\widehat{\boldsymbol{\beta}})^2$ is consistent for $\mathrm{Var}(\varepsilon)$. All that is left to show is that under the assumptions of Theorem 4.4.1 we have $\widehat{\mathbf{v}}_1$ is consistent for $\mathbf{v}^{*T}\boldsymbol{\Sigma}_{\boldsymbol{X}}\mathbf{v} = \mathbf{v}_1^*$. We have $|\mathbf{v}_1^* - \widehat{\mathbf{v}}_1| \leq \|\mathbf{v}^* - \mathbf{v}\|_1 = O_p(\lambda' s_{\mathbf{v}}) = o_p(1)$, by Lemma C.2.5. □

Proof of Proposition 4.4.8. Note that:

$$\widehat{\Delta}_3 = \underbrace{(\boldsymbol{\beta}^* - \widehat{\boldsymbol{\beta}})^T \underbrace{\frac{1}{n}\sum_{i=1}^n \boldsymbol{X}_i^{\otimes 2}\widehat{\mathbf{v}}^{\otimes 2}\boldsymbol{X}_i^{\otimes 2}}_{M}(\boldsymbol{\beta}^* - \widehat{\boldsymbol{\beta}})}_{I_1} + \underbrace{\frac{2}{n}\sum_{i=1}^n \widehat{\mathbf{v}}^T\boldsymbol{X}_i^{\otimes 2}(\boldsymbol{\beta}^* - \widehat{\boldsymbol{\beta}})\widehat{\mathbf{v}}^T\boldsymbol{X}_i\varepsilon_i}_{I_2} + \underbrace{\frac{1}{n}\sum_{i=1}^n (\widehat{\mathbf{v}}^T\boldsymbol{X}_i\varepsilon_i)^2}_{I_3}$$

We first handle I_1. We have that:

$$(\boldsymbol{\beta}^* - \widehat{\boldsymbol{\beta}})^T M(\boldsymbol{\beta}^* - \widehat{\boldsymbol{\beta}}) \leq \|\boldsymbol{\beta}^* - \widehat{\boldsymbol{\beta}}\|_1^2 \|M\|_\infty.$$

In much the same way as in the proof of Proposition 4.5.6 (see (C.3.1)), we can show that

$$\|M\|_\infty = O_p(\log(nd))\|\mathbf{v}^*\|_1.$$

Thus since $\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_1^2 = O_p\left(s^2\frac{\log d}{n}\right)$, we have

$$(\boldsymbol{\beta}^* - \widehat{\boldsymbol{\beta}})^T M(\boldsymbol{\beta}^* - \widehat{\boldsymbol{\beta}}) = O_p\left(s^2\frac{\log d}{n}\log(nd)\right)\|\mathbf{v}^*\|_1 = o_p(1),$$

by assumption.

Next, we take a look at I_3.

$$I_3 = \underbrace{\mathrm{Var}(\varepsilon)\frac{1}{n}\sum_{i=1}^{n}(\widehat{\mathbf{v}}^T\boldsymbol{X}_i)^2}_{I_{31}} + \underbrace{\frac{1}{n}\sum_{i=1}^{n}(\widehat{\mathbf{v}}^T\boldsymbol{X}_i)^2(\varepsilon_i^2 - \mathrm{Var}(\varepsilon))}_{I_{32}}.$$

Note that by the same proof as in Proposition 4.4.6 we can show that $I_{31} \to_P \Delta$. Now we show that I_{32} is small.

$$|I_{32}| \leq \|\widehat{\mathbf{v}}\|_1^2 \max_i \|\boldsymbol{X}_i^{\otimes 2}\|_{\max} \frac{1}{n}\sum_{i=1}^{n}(\varepsilon_i^2 - \mathrm{Var}(\varepsilon)).$$

By Lemma C.2.8 we have that $\max_i \|\boldsymbol{X}_i^{\otimes 2}\|_{\max} = O_p(\log(nd))$, and furthermore $\|\widehat{\mathbf{v}}\|_1^2 = \|\mathbf{v}^*\|_1^2 + o_p(1)$. Thus by Chebyshev's inequality we have:

$$|I_{32}| = O_p\left(\|\mathbf{v}^*\|_1^2 \frac{\log(nd)}{\sqrt{n}}\right) = o_p(1),$$

where the last follows by assumption. Finally, by Cauchy-Schwartz we have:

$$|I_2| \leq 2\sqrt{I_1}\sqrt{I_3} = o_p(1)O_p(1) = o_p(1),$$

which finishes the proof. □

Proof of Theorem 4.4.12. The proof of this Theorem follows from the general Theorem 4.3.22 upon verifying the assumptions. Note that in the proof of Theorem 4.4.10, we have verified all assumptions except for Assumption 4.3.21 and Assumption 4.3.18 in particular (4.3.14).

In the lines below we verify assumption 4.3.21. First observe that in the special case of a linear model $\mathbf{v}$ does not depend on the parameter $\boldsymbol{\beta}$. Therefore we will write $\mathbf{v}^*$. Note that we have the

following inequality:

$$\begin{aligned}\sqrt{n}|S(\theta,\boldsymbol{\gamma})-S(0,\boldsymbol{\gamma})-\theta| &= \sqrt{n}\theta\left|\frac{1}{n}\mathbf{v}^{*T}\left(\sum_{i=1}^{n}\boldsymbol{X}_i\boldsymbol{X}_{i,1}-\boldsymbol{\Sigma}_{\boldsymbol{X},*1}\right)\right| \\ &\leq \sqrt{n}\theta\|\mathbf{v}^*\|_1\underbrace{\left\|\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{X}_i\boldsymbol{X}_{i,1}-\boldsymbol{\Sigma}_{\boldsymbol{X},*1}\right\|_\infty}_{I}.\end{aligned}$$

Lemma C.2.2 shows that $|I| \leq \xi\sqrt{\frac{\log d}{n}}$ with high probability for a sufficiently large $\xi > 0$. Hence under our assumption since the RHS is independent of $\boldsymbol{\beta}$ we conclude that:

$$\lim_{n\to\infty}\inf_{\boldsymbol{\beta}\in\boldsymbol{\Omega}_1(K,\phi)}\mathbb{P}_{\boldsymbol{\beta}}\left(\sqrt{n}|S(\theta,\boldsymbol{\gamma})-S(0,\boldsymbol{\gamma})+\theta| \leq \xi\|\mathbf{v}^*\|_1 n^{-\phi}\sqrt{\log d}\right) = 1.$$

Next we verify (4.3.14). We have:

$$\left\|\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{X}_i(\boldsymbol{X}_{i,-1}^T\boldsymbol{\gamma}^*-Y_i)\right\|_\infty \leq \left\|\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{X}_i\varepsilon_i\right\|_\infty + Kn^{-\phi}\left\|\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{X}_i\boldsymbol{X}_{i,1}\right\|_\infty.$$

While we have a bound on the first term from (C.2.9) — $\left\|\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{X}_i\varepsilon_i\right\|_\infty \leq \xi'\sqrt{\frac{\log d}{n}}$ for some $\xi' > 0$, we don't immediately have a bound on the second term. Using the same idea as in Lemma C.2.2 we have that $\left\|\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{X}_i\boldsymbol{X}_{i,1}\right\|_\infty \leq \|\boldsymbol{\Sigma}_{\boldsymbol{X},*1}\|_\infty + C\sqrt{\frac{\log d}{n}} \leq 2K_{\boldsymbol{X}}^2 + C\sqrt{\frac{\log d}{n}}$, for some large constant C. Here we used the bound $\|\boldsymbol{\Sigma}_{\boldsymbol{X},*1}\|_\infty \leq 2K_{\boldsymbol{X}}^2$ which follows by the definition of ψ_2 norm.

This finishes the proof, since by Lemma C.2.5 we have:

$$\sqrt{n}\|\widehat{\mathbf{v}}-\mathbf{v}^*\|_1\left\|\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{X}_i(\boldsymbol{X}_{i,-1}^T\boldsymbol{\gamma}^*-Y_i)\right\|_\infty = O_p\left(\|\mathbf{v}^*\|_1 Kn^{-\phi}\sqrt{\log d}\right) = o_p(1).$$

□

In what follows we let $\mathbf{X}_{n\times d}$ be a matrix, which rows are the $\boldsymbol{X}_i^T$ vectors stacked together.

Lemma C.2.2. *We have that with probability at least* $1-2d^{2(1-c_X A_X^2)}$*:*

$$\left\|\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{X}_i^{\otimes 2}-\boldsymbol{\Sigma}_{\boldsymbol{X}}\right\|_{\max}=\|\boldsymbol{\Sigma}_n-\boldsymbol{\Sigma}_{\boldsymbol{X}}\|_{\max}\leq 2A_X K_{\boldsymbol{X}}^2\sqrt{\frac{\log d}{n}}.$$

Note. *The constant* c_X *is a universal constant independent of the* $\boldsymbol{X}$ *distribution,* $K_{\boldsymbol{X}}$ *is as defined in the main text, and* $A_X>0$ *is an arbitrarily chosen constant.*

Proof of Lemma C.2.2. First we note that the elements of the matrix – $\boldsymbol{X}^{\otimes 2}$ are sub-exponential random variables. This fact can be seen along the following lines. Note that for any fixed $p\geq 1$ and two univariate sub-Gaussian random variables X and Y, by the triangle inequality, and the simple inequity $a^2+b^2\geq 2|ab|$, we have:

$$\frac{[\mathbb{E}|2XY|^p]^{1/p}}{p}\leq 2\left(\frac{[\mathbb{E}|X|^{2p}]^{1/(2p)}}{\sqrt{2p}}\right)^2+2\left(\frac{[\mathbb{E}|Y|^{2p}]^{1/(2p)}}{\sqrt{2p}}\right)^2.$$

Hence:

$$\|\boldsymbol{X}^i\boldsymbol{X}^j\|_{\psi_1}\leq\|\boldsymbol{X}^i\|_{\psi_2}^2+\|\boldsymbol{X}^j\|_{\psi_2}^2\leq 2K_{\boldsymbol{X}}^2. \tag{C.2.4}$$

Using a Bernstein type of tail bound, for sub-exponential distributions (see Proposition 5.16 in Vershynin [84]) in addition to the union bound we get:

$$\mathbb{P}\left(\|\boldsymbol{\Sigma}_n-\boldsymbol{\Sigma}_{\boldsymbol{X}}\|_{\max}\geq t\right)\leq 2d^2\exp\left[-2c_X\min\left(\frac{t^2n}{4K_{\boldsymbol{X}}^4},\frac{tn}{2K_{\boldsymbol{X}}^2}\right)\right],$$

where c_X is a absolute constant independent of the distribution of X. Therefore plugging in $t=2A_XK_{\boldsymbol{X}}^2\sqrt{\frac{\log d}{n}}$, for a large enough constant A_X would yield that $\|\boldsymbol{\Sigma}_n-\boldsymbol{\Sigma}_{\boldsymbol{X}}\|_{\max}\leq 2A_XK_{\boldsymbol{X}}^2\sqrt{\frac{\log d}{n}}$ with probability at least $1-2d^{2-2c_XA_X^2}$, as claimed. □

Lemma C.2.3. *Assume the same conditions as in Lemma C.2.2, and assume further that the minimum eigenvalue* $\lambda_{\min}(\mathbf{\Sigma}_{\boldsymbol{X}}) > 0$ *and* $s\sqrt{\frac{\log d}{n}} \leq (1-\kappa)\frac{\lambda_{\min}(\mathbf{\Sigma}_{\boldsymbol{X}})}{(1+\xi)^2 2A_X K_{\boldsymbol{X}}^2}$*, where* $0 < \kappa < 1$*. We then have that* $\mathbf{\Sigma}_n$ *satisfies the RE property with* $\mathrm{RE}_{\mathbf{\Sigma}_n}(s,\xi) \geq \kappa\,\mathrm{RE}_{\mathbf{\Sigma}_{\boldsymbol{X}}}(s,\xi) \geq \kappa\lambda_{\min}(\mathbf{\Sigma}_{\boldsymbol{X}}) > 0$ *with probability at least* $1 - 2d^{2-2c_X A_X^2}$.

Proof of Lemma C.2.3. Take a non-zero vector in the cone: $\mathbf{u} \in \{\|\mathbf{u}_{S^c}\|_1 \leq \xi\|\mathbf{u}_S\|_1\}$, with $|S| \leq s$. Note that we have the following:

$$\begin{aligned}\left|\mathbf{u}^T\mathbf{\Sigma}_n\mathbf{u} - \mathbf{u}^T\mathbf{\Sigma}_{\boldsymbol{X}}\mathbf{u}\right| &\leq \|\mathbf{u}\|_1^2 \left\|\mathbf{\Sigma}_n - \mathbf{\Sigma}_{\boldsymbol{X}}\right\|_{\max} \\ &\leq (1+\xi)^2\|\mathbf{u}_S\|_1^2 \left\|\mathbf{\Sigma}_n - \mathbf{\Sigma}_{\boldsymbol{X}}\right\|_{\max} \\ &\leq (1+\xi)^2 s\|\mathbf{u}_S\|_2^2 \left\|\mathbf{\Sigma}_n - \mathbf{\Sigma}_{\boldsymbol{X}}\right\|_{\max}.\end{aligned}$$

The last of course implies:

$$\mathrm{RE}_{\mathbf{\Sigma}_n}(s,\xi) \geq \mathrm{RE}_{\mathbf{\Sigma}_{\boldsymbol{X}}}(s,\xi) - s(1+\xi)^2 \left\|\mathbf{\Sigma}_n - \mathbf{\Sigma}_{\boldsymbol{X}}\right\|_{\max}.$$

Now, on the event:

$$\left\|\mathbf{\Sigma}_n - \mathbf{\Sigma}_{\boldsymbol{X}}\right\|_{\max} \leq 2A_X K_{\boldsymbol{X}}^2\sqrt{\frac{\log d}{n}},$$

we have:

$$\mathrm{RE}_{\mathbf{\Sigma}_n}(s,\xi) \geq \mathrm{RE}_{\mathbf{\Sigma}_{\boldsymbol{X}}}(s,\xi) - s(1+\xi)^2 2A_X K_{\boldsymbol{X}}^2\sqrt{\frac{\log d}{n}}.$$

Thus if $s\sqrt{\frac{\log d}{n}} \leq (1-\kappa)\frac{\lambda_{\min}(\mathbf{\Sigma}_{\boldsymbol{X}})}{(1+\xi)^2 2A_X K_{\boldsymbol{X}}^2}$, for some $0 < \kappa < 1$ we conclude that:

$$\mathrm{RE}_{\mathbf{\Sigma}_n}(s,\xi) \geq \kappa\,\mathrm{RE}_{\mathbf{\Sigma}_{\boldsymbol{X}}}(s,\xi) \geq \kappa\lambda_{\min}(\mathbf{\Sigma}_{\boldsymbol{X}}) > 0,$$

where the probability bound on the event follows from Lemma C.2.2.

□

Definition C.2.4. *For a fixed* $0 < \kappa < 1$*, let* $\mathrm{RE}_\kappa(s, \xi) = \kappa \, \mathrm{RE}_{\boldsymbol{\Sigma_X}}(s, \xi)$.

Lemma C.2.5. *Assume that —* $\lambda_{\min}(\boldsymbol{\Sigma_X}) > \delta > 0$*,* $s_{\mathbf{v}} \sqrt{\frac{\log d}{n}} \leq (1 - \kappa) \frac{\lambda_{\min}(\boldsymbol{\Sigma_X})}{(1+1)^2 2 A_X K_{\boldsymbol{X}}^2}$*, where* $0 < \kappa < 1$ *and* $\lambda' \geq \|\mathbf{v}^*\|_1 2 A_X K_{\boldsymbol{X}}^2 \sqrt{\frac{\log d}{n}}$*. Then we have that* $\|\widehat{\mathbf{v}} - \mathbf{v}^*\|_1 \leq \frac{8 \lambda' s_{\mathbf{v}}}{\mathrm{RE}_\kappa(s_{\mathbf{v}}, 1)}$ *with probability at least* $1 - 2d^{2 - 2 c_X A_X^2}$.

Proof of Lemma C.2.5. We start by showing that $\mathbf{v}^*$ satisfies the Dantzig selector constraint, i.e.

$$\left\| \frac{1}{n} \sum_{i=1}^n \mathbf{v}^{*T} \boldsymbol{X}_i^{\otimes 2} - \mathbf{e} \right\|_\infty \leq \lambda',$$

with high probability. To this end, note that:

$$\begin{aligned} \left\| \frac{1}{n} \sum_{i=1}^n \mathbf{v}^{*T} \boldsymbol{X}_i^{\otimes 2} - \mathbf{e} \right\|_\infty &\leq \|\mathbf{v}^*\|_1 \left\| \boldsymbol{\Sigma}_n - \boldsymbol{\Sigma_X} \right\|_{\max} \\ &\leq \|\mathbf{v}^*\|_1 2 A_X K_{\boldsymbol{X}}^2 \sqrt{\frac{\log d}{n}}, \end{aligned}$$

where the last inequality holds with probability at least $1 - 2d^{2 - 2 c_X A_X^2}$ as in Lemma C.2.2. Thus for values of $\lambda' \geq \|\mathbf{v}^*\|_1 2 A_X K_{\boldsymbol{X}}^2 \sqrt{\frac{\log d}{n}}$, the above gives us that:

$$\begin{aligned} \left\| \frac{1}{n} \sum_{i=1}^n (\widehat{\mathbf{v}} - \mathbf{v}^*)^T \boldsymbol{X}_i^{\otimes 2} \right\|_\infty &\leq \left\| \frac{1}{n} \sum_{i=1}^n \widehat{\mathbf{v}}^T \boldsymbol{X}_i^{\otimes 2} - \mathbf{e} \right\|_\infty + \left\| \frac{1}{n} \sum_{i=1}^n \mathbf{v}^{*T} \boldsymbol{X}_i^{\otimes 2} - \mathbf{e} \right\|_\infty \\ &\leq 2\lambda'. \end{aligned} \tag{C.2.5}$$

Let $S_{\mathbf{v}} = \mathrm{supp}(\mathbf{v}^*)$, with $s_{\mathbf{v}} = |S_{\mathbf{v}}|$. We can therefore conclude that:

$$\|\mathbf{v}^*_{S_{\mathbf{v}}}\|_1 = \|\mathbf{v}^*\|_1 \geq \|\widehat{\mathbf{v}}\|_1 = \|\widehat{\mathbf{v}}_{S_{\mathbf{v}}}\|_1 + \|\widehat{\mathbf{v}}_{S_{\mathbf{v}}^c}\|_1.$$

Furthermore, by the triangle inequality:

$$\|\widehat{\mathbf{v}}_{S_{\mathbf{v}}}\|_1 \geq \|\mathbf{v}^*_{S_{\mathbf{v}}}\|_1 - \|\widehat{\mathbf{v}}_{S_{\mathbf{v}}} - \mathbf{v}^*_{S_{\mathbf{v}}}\|_1.$$

Combining the last two inequalities we get that:

$$\|\widehat{\mathbf{v}}_{S^c_{\mathbf{v}}} - \mathbf{v}^*_{S^c_{\mathbf{v}}}\|_1 \leq \|\widehat{\mathbf{v}}_{S_{\mathbf{v}}} - \mathbf{v}^*_{S_{\mathbf{v}}}\|_1. \tag{C.2.6}$$

Now we evaluate:

$$\begin{aligned}\frac{1}{n}\|\mathbf{X}(\widehat{\mathbf{v}} - \mathbf{v}^*)\|_2^2 &\leq \|\boldsymbol{\Sigma}_n(\widehat{\mathbf{v}} - \mathbf{v}^*)\|_\infty \|(\widehat{\mathbf{v}} - \mathbf{v}^*)\|_1 \\ &\overset{\text{by (C.2.5),(C.2.6)}}{\leq} 2\lambda'(2\|\widehat{\mathbf{v}}_{S_{\mathbf{v}}} - \mathbf{v}^*_{S_{\mathbf{v}}}\|_1) \\ &\leq 4\lambda'\sqrt{s_{\mathbf{v}}}\|\widehat{\mathbf{v}}_{S_{\mathbf{v}}} - \mathbf{v}^*_{S_{\mathbf{v}}}\|_2.\end{aligned}$$

On the other hand, by Lemma C.2.3, we know that the matrix $\boldsymbol{\Sigma}_n$ satisfies the RE condition with $\mathrm{RE}_\kappa(s_{\mathbf{v}}, 1)$, on the same event on which we are working on, provided that $s_{\mathbf{v}}\sqrt{\frac{\log d}{n}} \leq (1 - \kappa)\frac{\lambda_{\min}(\boldsymbol{\Sigma}_{\boldsymbol{X}})}{(1+1)^2 2A_X K^2_{\boldsymbol{X}}}$, for some $\kappa < 1$. This implies that:

$$\frac{1}{n}\|\mathbf{X}(\widehat{\mathbf{v}} - \mathbf{v}^*)\|_2^2 \geq \mathrm{RE}_\kappa(s_{\mathbf{v}}, 1)\|\widehat{\mathbf{v}}_{S_{\mathbf{v}}} - \mathbf{v}^*_{S_{\mathbf{v}}}\|_2^2.$$

The last inequality gives us that:

$$\|\widehat{\mathbf{v}} - \mathbf{v}^*\|_1 \leq 2\|\widehat{\mathbf{v}}_{S_{\mathbf{v}}} - \mathbf{v}^*_{S_{\mathbf{v}}}\|_1 \leq 2\sqrt{s_{\mathbf{v}}}\|\widehat{\mathbf{v}}_{S_{\mathbf{v}}} - \mathbf{v}^*_{S_{\mathbf{v}}}\|_2 \leq \frac{8\lambda' s_{\mathbf{v}}}{\mathrm{RE}_\kappa(s_{\mathbf{v}}, 1)},$$

with probability at least $1 - 2d^{2-2c_X A_X^2}$, as claimed. □

Lemma C.2.6. *Assume the same conditions as in Lemma C.2.2 and that* $\sqrt{\frac{\log d}{n}} \leq C$ *for some*

constant C. Let $S_0 = \operatorname{supp}(\boldsymbol{\beta}^*)$, and let $\lambda = AK\sqrt{\frac{\log d}{n}}$. Then, with probability at least $1 - ed^{1-\frac{cA^2}{2(1+CA_X)K_{\boldsymbol{X}}^2}} - 2d^{2(1-c_X A_X^2)}$ (where c is a universal constant independent of the distribution of ε, $K = \|\varepsilon\|_{\psi_2}$, and the other constants are defined in Lemma C.2.2) we have:

$$\|\widehat{\boldsymbol{\beta}}_{S_0^c} - \boldsymbol{\beta}^*_{S_0^c}\|_1 \leq \|\widehat{\boldsymbol{\beta}}_{S_0} - \boldsymbol{\beta}^*_{S_0}\|_1, \tag{C.2.7}$$

and:

$$\left\|\boldsymbol{\Sigma}_n(\boldsymbol{\beta}^* - \widehat{\boldsymbol{\beta}})\right\|_\infty \leq 2\lambda. \tag{C.2.8}$$

Proof of Lemma C.2.6. Note that by a Hoeffding's type of inequality for sub-Gaussian random variables (see Proposition 5.10[84]) and the union bound, we have:

$$\mathbb{P}\left(\left\|\frac{1}{n}\mathbf{X}^T\boldsymbol{\varepsilon}\right\|_\infty \geq t \middle| \mathbf{X}\right) \leq ed\exp\left(-\frac{cnt^2}{K^2\|\boldsymbol{\Sigma}_n\|_\infty}\right), \tag{C.2.9}$$

where c is a universal constant. Under the assumption that $\sqrt{\frac{\log d}{n}} < C$, we have that on the event considered in Lemma C.2.2, that $\|\boldsymbol{\Sigma}_n\|_{\max} \leq \|\boldsymbol{\Sigma}_{\boldsymbol{X}}\|_{\max} + 2CA_X K_{\boldsymbol{X}}^2$ with probability at least $1 - 2d^{2(1-c_X A_X^2)}$. Note that $\|\boldsymbol{\Sigma}_{\boldsymbol{X}}\|_{\max} \leq \max_{i=1,\ldots,d}\mathbb{E}(\boldsymbol{X}^i)^2 \leq 2K_{\boldsymbol{X}}^2$, by the sub-Gaussian assumption on $\boldsymbol{X}$ and the definition of ψ_2 norm. Hence $\|\boldsymbol{\Sigma}_n\|_{\max} \leq 2(1 + CA_X)K_{\boldsymbol{X}}^2$. Let $E_X = \{\|\boldsymbol{\Sigma}_n\|_{\max} \leq 2(1 + CA_X)K_{\boldsymbol{X}}^2\}$.

Thus on the event E_X, setting the value $t = \lambda = AK\sqrt{\frac{\log d}{n}}$, the probability bound (C.2.9) becomes $ed^{1-\frac{cA^2}{2(1+CA_X)K_{\boldsymbol{X}}^2}}$. Denote with $E = \left\{\left\|\frac{1}{n}\mathbf{X}^T\boldsymbol{\varepsilon}\right\|_\infty \leq \lambda\right\} \cap E_X$, which holds with probability at least $1 - ed^{1-\frac{cA^2}{2(1+CA_X)K_{\boldsymbol{X}}^2}} - 2d^{2(1-c_X A_X^2)}$ by the union bound.

Note that when E holds, the true parameter satisfies the Dantzig selector constraint and thus

we can obtain (C.2.7) in the same manner as in Lemma C.2.5. To obtain (C.2.8), note that by the triangle inequality on the event E we have:

$$\begin{aligned}\left\|\boldsymbol{\Sigma}_n(\boldsymbol{\beta}^* - \widehat{\boldsymbol{\beta}})\right\|_\infty &\leq \left\|\frac{1}{n}\mathbf{X}^T\boldsymbol{\varepsilon}\right\|_\infty + \left\|\frac{1}{n}\mathbf{X}^T(\boldsymbol{Y} - \mathbf{X}\widehat{\boldsymbol{\beta}})\right\|_\infty \\ &\leq 2\lambda,\end{aligned}$$

with probability at least $1 - ed^{1-\frac{cA^2}{2(1+CA_X)K_{\boldsymbol{X}}^2}} - 2d^{2(1-c_X A_X^2)}$ as claimed. □

Lemma C.2.7. *Assume the same conditions in Lemmas C.2.2, C.2.3 (with $\xi = 1$), and C.2.6, so that $\boldsymbol{\Sigma}_n$ satisfies the RE assumption with $\mathrm{RE}_\kappa(s, 1)$ with high probability. Set $\lambda = AK\sqrt{\frac{\log d}{n}}$, as in Lemma C.2.6. Then with probability at least $1 - ed^{1-\frac{cA^2}{2(1+CA_X)K_{\boldsymbol{X}}^2}} - 2d^{2(1-c_X A_X^2)}$ we have:*

$$\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_1 \leq \frac{8AK}{\mathrm{RE}_\kappa(s,1)} s\sqrt{\frac{\log d}{n}}, \tag{C.2.10}$$

$$\|\mathbf{X}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)\|_2^2 \leq \frac{16A^2K^2}{\mathrm{RE}_\kappa(s,1)} s\log d. \tag{C.2.11}$$

Proof. From Lemma C.2.6, we know that on the event E (which happens with probability at least $1 - ed^{1-\frac{cA^2}{2(1+CA_X)K_{\boldsymbol{X}}^2}} - 2d^{2(1-c_X A_X^2)}$) we have, that (C.2.7) and (C.2.8) hold. Thus on the event E, we have:

$$\begin{aligned}\frac{1}{n}\|\mathbf{X}(\boldsymbol{\beta}^* - \widehat{\boldsymbol{\beta}})\|_2^2 &\leq \|\boldsymbol{\Sigma}_n(\boldsymbol{\beta}^* - \widehat{\boldsymbol{\beta}})\|_\infty \|(\boldsymbol{\beta}^* - \widehat{\boldsymbol{\beta}})\|_1 \\ &\leq 2\lambda(2\|\boldsymbol{\beta}^*_{S_0} - \widehat{\boldsymbol{\beta}}_{S_0}\|_1) \\ &\leq 4\lambda\sqrt{s}\|\boldsymbol{\beta}^*_{S_0} - \widehat{\boldsymbol{\beta}}_{S_0}\|_2.\end{aligned}$$

Now note that on the event E — $\frac{1}{n}\mathbf{X}^T\mathbf{X} = \mathbf{\Sigma}_n$ satisfies the RE condition and thus we have:

$$\frac{1}{n}\|\mathbf{X}(\boldsymbol{\beta}^* - \widehat{\boldsymbol{\beta}})\|_2^2 \geq \mathrm{RE}_\kappa(s,1)\|\boldsymbol{\beta}^*_{S_0} - \widehat{\boldsymbol{\beta}}_{S_0}\|_2^2.$$

The last two inequalities yield:

$$\|\boldsymbol{\beta}^*_{S_0} - \widehat{\boldsymbol{\beta}}_{S_0}\|_2 \leq \frac{4\lambda\sqrt{s}}{\mathrm{RE}_\kappa(s,1)},$$
$$\frac{1}{n}\|\mathbf{X}(\boldsymbol{\beta}^* - \widehat{\boldsymbol{\beta}})\|_2^2 \leq \frac{16\lambda^2 s}{\mathrm{RE}_\kappa(s,1)}.$$

The last inequality actually gives (C.2.11). To get (C.2.10), note that by (C.2.7) we have:

$$\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_1 \leq 2\|\widehat{\boldsymbol{\beta}}_{S_0} - \boldsymbol{\beta}^*_{S_0}\|_1 \leq \frac{8AK\lambda s}{\mathrm{RE}_\kappa(s,1)},$$

and we are done. □

Lemma C.2.8. *Let $\{\boldsymbol{X}_i\}_{i=1}^n$ are identical (not necessarily independent), d-dimensional sub-Gaussian vectors with $\max_{\substack{i=1,\dots,n\\ j=1,\dots,d}} \|\boldsymbol{X}_i^j\|_{\psi_2} = K$. Then we have:*

$$\max_{i=1,\dots,n} \|\boldsymbol{X}_i^{\otimes 2}\|_{\max} = O_p(\log(nd)).$$

Proof of Lemma C.2.8. Note that by the union bound for a fixed i we have:

$$\mathbb{P}(\|\boldsymbol{X}_i\|_\infty \geq t) \leq d\exp(1 - ct^2/K^2),$$

where c is an absolute constant, by (5.10) in Vershynin [84]. With yet another union bound we get:

$$\mathbb{P}(\max_{i=1,\dots,n} \|\boldsymbol{X}_i\|_\infty \geq t) \leq nd\exp(1 - ct^2/K^2).$$

Thus as long as $t \geq C\sqrt{\log(nd)}$ for a large enough C the above probability will converge to o. This finishes the proof, as clearly:

$$\max_{i=1,\ldots,n} \|\boldsymbol{X}_i^{\otimes 2}\|_{\max} \leq \max_{i=1,\ldots,n} \|\boldsymbol{X}_i\|_\infty^2 \leq t^2.$$

□

Lemma C.2.9. *Let $R \subset \{1,\ldots,d\}$ with $|R| = r$. Then we have the following:*

$$\mathbb{E}\|(\boldsymbol{X}\varepsilon)_R\|_2^3 \leq r^{3/2}(6KK_{\boldsymbol{X}})^3.$$

Proof. First we use Jensen's inequality to get:

$$\mathbb{E}\|(\boldsymbol{X}\varepsilon)_R\|_2^3 \leq \sqrt{\mathbb{E}\varepsilon^6}\sqrt{\mathbb{E}\left(\sum_{j\in R}\boldsymbol{X}^{j2}\right)^3}.$$

The first term is clearly bounded $\sqrt{\mathbb{E}\varepsilon^6} \leq (\sqrt{6}K)^3$, by the definition of K. To deal with the second term, first note:

$$\left(\frac{\sum_{j\in R}\boldsymbol{X}^{j2}}{r}\right)^3 \leq \left(\frac{\sum_{j\in R}|\boldsymbol{X}^j|^3}{r}\right)^2,$$

which follows from the generalized mean inequality (or monotonicity of the L_p norms). Thus by the triangle inequality we have:

$$\begin{aligned}\sqrt{\mathbb{E}\left(\frac{\sum_{j\in R}\boldsymbol{X}^{j2}}{r}\right)^3} \leq \sqrt{\mathbb{E}\left(\frac{\sum_{j\in R}|\boldsymbol{X}^j|^3}{r}\right)^2} &\leq \sum_{j\in R}\sqrt{\mathbb{E}\left(\frac{|\boldsymbol{X}^j|^6}{r^2}\right)} \\ &\leq (\sqrt{6}K_{\boldsymbol{X}})^3. \end{aligned} \tag{C.2.12}$$

Hence

$$\mathbb{E}\|(\boldsymbol{X}\varepsilon)_R\|_2^3 \leq r^{3/2}(6KK_{\boldsymbol{X}})^3,$$

as claimed. □

C.3 Proofs for Edge Testing

C.3.1 Proofs for CLIME

Proof of Theorem 4.5.2. To show this theorem, we simply need to verify the conditions of Theorem 4.3.3. Using Lemma C.2.5, we have that $\|\widehat{\mathbf{v}} - \mathbf{v}^*\|_1 = O_p\left(\|\mathbf{v}^*\|_1 s_{\mathbf{v}}\sqrt{\frac{\log d}{n}}\right)$, provided that $\lambda' \asymp \|\mathbf{v}^*\|_1\sqrt{\frac{\log d}{n}}$ is large enough (see the Lemma for details). Note that Lemma C.2.5, also shows that the term $\left\|\frac{1}{n}\sum_{i=1}^n \boldsymbol{X}_i\boldsymbol{X}_{i,-1}^T\boldsymbol{\gamma}^* - \mathbf{e}_m^T\right\|_\infty \leq \lambda$ with high probability, provided that $\lambda \asymp \|\boldsymbol{\gamma}^*\|_1\sqrt{\frac{\log d}{n}}$ (note that $\|\boldsymbol{\gamma}^*\|_1 = \|\boldsymbol{\beta}^*\|_1$) is large enough.

Hence, by assumption we have:

$$n^{1/2}\left\|\frac{1}{n}\sum_{i=1}^n \boldsymbol{X}_i\boldsymbol{X}_{i,-j}^T\boldsymbol{\gamma}^* - \mathbf{e}_m^T\right\|_\infty \|\widehat{\mathbf{v}} - \mathbf{v}^*\|_1 = O_p\left(\|\boldsymbol{\gamma}^*\|_1\|\mathbf{v}^*\|_1 s_{\mathbf{v}}\frac{\log d}{\sqrt{n}}\right) = o_p(1).$$

Moreover Lemma C.2.5 shows that it sufficicient to select $\lambda' \asymp \|\mathbf{v}^*\|_1\sqrt{\frac{\log d}{n}}$, and again by Lemma C.2.5 we know that $\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_1 = O_p\left(\|\boldsymbol{\gamma}^*\|_1 s\sqrt{\frac{\log d}{n}}\right)$. Hence in our assumed regime:

$$\max(s_{\mathbf{v}}, s)\|\mathbf{v}^*\|_1\|\boldsymbol{\gamma}^*\|_1\frac{\log d}{\sqrt{n}} = o(1),$$

we have:

$$n^{1/2}\lambda'\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_1 = o_p(1),$$

which is condition (4.3.5) from Theorem 4.3.3, and finishes the proof. □

Proof of Corollary 4.5.4. Similarly to Corollary 4.4.5 we will verify Lyapunov's condition for the CLT. It suffices to bound the quantity:

$$\frac{n^{-3/2}}{\|\mathbf{v}^*\|_2^3\|\boldsymbol{\beta}^*\|_2^3\iota_{\min}^{3/2}}\sum_{i=1}^n \mathbb{E}\left|\mathbf{v}^{*T}\boldsymbol{X}_i^{\otimes 2}\boldsymbol{\beta}^* - \mathbf{v}^{*T}\boldsymbol{\Sigma}_{\boldsymbol{X}}\boldsymbol{\beta}^*\right|^3,$$

where we used assumption (4.5.2). By Cauchy-Schwartz we can bound the above expression by following (up to a constant factor):

$$n^{-3/2}\sum_{i=1}^n \mathbb{E}\left\|\mathrm{Vec}[(\boldsymbol{X}_i^{\otimes 2} - \boldsymbol{\Sigma}_{\boldsymbol{X}})_{S_{\mathbf{v}},S}]\right\|_2^3,$$

where by subscripting the matrix we mean setting all elements not in the supports of $\mathbf{v}^*$ or $\boldsymbol{\beta}^*$ ($S_{\mathbf{v}}$, and S correspondigly) to 0, and the operator Vec vectorizes the matrix. Finally using Lemma C.3.2 we conclude that we can control the expression above by:

$$\frac{(s_{\mathbf{v}}s)^{3/2}}{n^{1/2}(24K_{\boldsymbol{X}}^2)^3},$$

and hence the conclusion follows. □

Remark C.3.1. *Using the Berry-Esseen theorem for non-identical random variables in combination with the bound we derived above, we can further show:*

$$\sup_t\left|\mathbb{P}^*\left(\frac{n^{1/2}}{\sqrt{\mathrm{Var}(\mathbf{v}^{*T}\boldsymbol{X}^{\otimes 2}\boldsymbol{\beta}^*)}}S(0,\boldsymbol{\gamma}^*)\le t\right) - \Phi(t)\right| \le C_{BE}\frac{n^{-1/2}}{(24K_{\boldsymbol{X}}^2)^3}(s_{\mathbf{v}}s)^{3/2} = o(1),$$

where C_{BE} is an absolute constant.

Proof of Proposition 4.5.6. Note that:

$$\frac{1}{n}\sum_{i=1}^n(\widehat{\mathbf{v}}^T(\boldsymbol{X}_i^{\otimes 2} - \boldsymbol{\Sigma}_n)\widehat{\boldsymbol{\beta}})^2 = \frac{1}{n}\sum_{i=1}^n(\widehat{\mathbf{v}}^T\boldsymbol{X}_i^{\otimes 2}\widehat{\boldsymbol{\beta}})^2 - (\widehat{\mathbf{v}}^T\boldsymbol{\Sigma}_n\widehat{\boldsymbol{\beta}})^2.$$

First we show that $\widehat{\mathbf{v}}^T \boldsymbol{\Sigma}_n \widehat{\boldsymbol{\beta}}$ is consistent for $\mathbf{v}^{*T} \boldsymbol{\Sigma}_{\boldsymbol{X}} \boldsymbol{\beta}^*$. We have:

$$|\widehat{\mathbf{v}}^T \boldsymbol{\Sigma}_n \widehat{\boldsymbol{\beta}} - \mathbf{v}^{*T} \boldsymbol{\Sigma}_n \boldsymbol{\beta}^*| \leq |\underbrace{\widehat{\mathbf{v}}^T \boldsymbol{\Sigma}_n \widehat{\boldsymbol{\beta}} - \mathbf{v}^{*T} \boldsymbol{\Sigma}_n \widehat{\boldsymbol{\beta}}}_{I_1}| + |\underbrace{\mathbf{v}^{*T} \boldsymbol{\Sigma}_n \widehat{\boldsymbol{\beta}} - \mathbf{v}^{*T} \boldsymbol{\Sigma}_n \boldsymbol{\beta}^*}_{I_2}|.$$

For I_1 we have:

$$|I_1| \leq \|\widehat{\mathbf{v}} - \mathbf{v}^*\|_1 \|\boldsymbol{\Sigma}_n \widehat{\boldsymbol{\beta}}\|_\infty \leq \|\widehat{\mathbf{v}} - \mathbf{v}^*\|_1 (1 + \lambda),$$

using Lemma C.2.5 in the appendix, we have that $\|\widehat{\mathbf{v}} - \mathbf{v}^*\|_1 = O_p\left(\|\mathbf{v}^*\|_1 s_{\mathbf{v}} \sqrt{\frac{\log d}{n}}\right)$, provided that $\lambda' \asymp \|\mathbf{v}^*\|_1 \sqrt{\frac{\log d}{n}}$ is large enough (see the Lemma for details), and since we are also taking $\lambda = O_p\left(\|\boldsymbol{\beta}^*\|_1 \sqrt{\frac{\log d}{n}}\right)$, we have that $|I_1| = o_p(1)$, by our assumption. Similarly for I_2 we have:

$$|I_2| \leq \|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_1 \|\mathbf{v}^{*T} \boldsymbol{\Sigma}_n\| = O_p\left(\|\boldsymbol{\beta}^*\|_1 s \sqrt{\frac{\log d}{n}}\right) O_p(1 + \lambda') = o(1),$$

where we used Lemma C.2.5 again. This completes the proof of the consistency of the plugin expectation estimator.

Next, we tackle the second moment plugin estimator:

$$\left|\frac{1}{n}\sum_{i=1}^n (\widehat{\mathbf{v}}^T \boldsymbol{X}_i^{\otimes 2} \widehat{\boldsymbol{\beta}})^2 - \mathbb{E}(\mathbf{v}^{*T} X^{\otimes 2} \boldsymbol{\beta}^*)^2\right| \leq \underbrace{\left|\frac{1}{n}\sum_{i=1}^n [(\widehat{\mathbf{v}}^T \boldsymbol{X}_i^{\otimes 2} \widehat{\boldsymbol{\beta}})^2 - (\mathbf{v}^{*T} \boldsymbol{X}_i^{\otimes 2} \boldsymbol{\beta}^*)^2]\right|}_{I_3}$$
$$+ \underbrace{\left|\frac{1}{n}\sum_{i=1}^n (\mathbf{v}^{*T} \boldsymbol{X}_i^{\otimes 2} \boldsymbol{\beta}^*)^2 - \mathbb{E}(\mathbf{v}^{*T} \boldsymbol{X}^{\otimes 2} \boldsymbol{\beta}^*)^2\right|}_{I_4}.$$

$$|I_3| \leq \underbrace{\left|\frac{1}{n}\sum_{i=1}^n [(\widehat{\mathbf{v}}^T \boldsymbol{X}_i^{\otimes 2} \widehat{\boldsymbol{\beta}})^2 - (\mathbf{v}^{*T} \boldsymbol{X}_i^{\otimes 2} \widehat{\boldsymbol{\beta}})^2]\right|}_{I_{31}} + \underbrace{\left|\frac{1}{n}\sum_{i=1}^n [(\mathbf{v}^{*T} \boldsymbol{X}_i^{\otimes 2} \widehat{\boldsymbol{\beta}})^2 - (\mathbf{v}^{*T} \boldsymbol{X}_i^{\otimes 2} \boldsymbol{\beta}^*)^2]\right|}_{I_{32}}.$$

For the first term we have:

$$I_{31} = (\widehat{\mathbf{v}} - \mathbf{v}^*)^T \underbrace{\frac{1}{n}\sum_{i=1}^{n} \boldsymbol{X}_i^{\otimes 2}\widehat{\boldsymbol{\beta}}\widehat{\boldsymbol{\beta}}^T\boldsymbol{X}_i^{\otimes 2}}_{M}(\widehat{\mathbf{v}} + \mathbf{v}^*).$$

Using Lemma C.2.8, we can handle M in the following way:

$$\|M\|_{\max} \leq \max \|\boldsymbol{X}_i^{\otimes 2}\|_{\max}\frac{1}{n}\sum_{i=1}^{n}\widehat{\boldsymbol{\beta}}^T\boldsymbol{X}_i^{\otimes 2}\widehat{\boldsymbol{\beta}} \leq O_p(\log(nd))\|\widehat{\boldsymbol{\beta}}\|_1\|\widehat{\boldsymbol{\Sigma}}_n\widehat{\boldsymbol{\beta}}\|_\infty. \tag{C.3.1}$$

By the definition of $\widehat{\boldsymbol{\beta}}$ we have: $\|\boldsymbol{\Sigma}_n\widehat{\boldsymbol{\beta}}\|_\infty \leq (1+\lambda)$. Hence:

$$\begin{aligned}\|M\|_{\max} &\leq O_p(\log(nd))(\|\boldsymbol{\beta}^*\|_1 + \|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_1)(1+\lambda)\\ &= O_p(\log(nd))\|\boldsymbol{\beta}^*\|_1,\end{aligned}$$

where we used that $\lambda \asymp \|\boldsymbol{\beta}^*\|_1\sqrt{\frac{\log d}{n}}$ and $\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_1 = O_p\left(\|\boldsymbol{\beta}^*\|_1 s\sqrt{\frac{\log d}{n}}\right)$, which are quantities going to 0, under our assumptions and furthermore $\|\boldsymbol{\beta}^*\|_1 \geq (2K_{\boldsymbol{X}}^2)^{-1} > 0$. Thus:

$$|I_{31}| \leq (\|\widehat{\mathbf{v}} - \mathbf{v}^*\|_1^2 + 2\|\widehat{\mathbf{v}} - \mathbf{v}^*\|_1\|\mathbf{v}^*\|_1)O_p(\log(nd))\|\boldsymbol{\beta}^*\|_1.$$

By Lemma C.2.5, $\|\widehat{\mathbf{v}} - \mathbf{v}^*\|_1 = O_p\left(s_{\mathbf{v}}\|\mathbf{v}^*\|_1\sqrt{\frac{\log d}{n}}\right)$, and since $\|\widehat{\mathbf{v}} - \mathbf{v}^*\|_1 = o(1)$, we have that:

$$|I_{31}| = O_p\left(s_{\mathbf{v}}\|\mathbf{v}^*\|_1^2\|\boldsymbol{\beta}^*\|_1\sqrt{\frac{\log d}{n}}\log(nd)\right) = o_p(1),$$

by assumption. By a similar argument we can show that I_{32} is $o_p(1)$. Finally we show that I_4 is

small. By Chebyshev's inequality and the finite variance assumption, we have:

$$I_4 = \left[\frac{1}{n}\sum_{i=1}^{n}(\mathbf{v}^{*T}\boldsymbol{X}_i^{\otimes 2}\boldsymbol{\beta}^*)^2 - \mathbb{E}(\mathbf{v}^{*T}\boldsymbol{X}^{\otimes 2}\boldsymbol{\beta}^*)^2\right] = O_p\left(\frac{\mathrm{Var}((\mathbf{v}^{*T}\boldsymbol{X}^{\otimes 2}\boldsymbol{\beta}^*)^2)}{n}\right) = o_p(1).$$

This completes the proof. □

Proof of Remark 4.5.7. We have that:

$$\frac{1}{n}\sum_{i=1}^{n}(\widehat{\mathbf{v}}^T\boldsymbol{X}_i^{\otimes 2}\widehat{\boldsymbol{\beta}} - \widehat{\mathbf{v}}^T\mathbf{e}_m^T)^2 = \underbrace{\frac{1}{n}\sum_{i=1}^{n}(\widehat{\mathbf{v}}^T\boldsymbol{X}_i^{\otimes 2}\widehat{\boldsymbol{\beta}})^2}_{I_1} - 2\underbrace{(\widehat{\mathbf{v}}^T\boldsymbol{\Sigma}_n\widehat{\boldsymbol{\beta}})\widehat{\mathbf{v}}^T\mathbf{e}_m^T}_{I_2} + \underbrace{(\widehat{\mathbf{v}}^T\mathbf{e}_m^T)^2}_{I_3}.$$

As a consequence of the proof of Proposition 4.5.6, we have that $I_1 \to_p \mathbb{E}(\mathbf{v}^{*T}\boldsymbol{X}^{\otimes 2}\boldsymbol{\beta}^*)^2$, also that $\widehat{\mathbf{v}}^T\boldsymbol{\Sigma}_n\widehat{\boldsymbol{\beta}} \to_p \mathbf{v}^{*T}\boldsymbol{\Sigma}_{\boldsymbol{X}}\boldsymbol{\beta}^*$. Thus, with the help of the continuous mapping theorem, all it remains to show is that: $\widehat{\mathbf{v}}^T\mathbf{e}_m^T$ is consistent for $\mathbf{v}^{*T}\mathbf{e}_m^T$. However this follows from:

$$|\widehat{\mathbf{v}}^T\mathbf{e}_m^T - \mathbf{v}^{*T}\mathbf{e}_m^T| \leq \|\widehat{\mathbf{v}} - \mathbf{v}^*\|_1 = o_p(1),$$

by Lemma C.2.5. This completes the proof. □

Proof of Theorem 4.5.11. To prove this theorem note that all bounds we showed in the proof of Theorem 4.5.9 hold uniformly in the parameter set $\mathcal{S}_0(L, s)$. Note that as both $\mathbf{v}$ and $\boldsymbol{\beta}$ are columns of $\boldsymbol{\Omega}$ we have that $\|\mathbf{v}\|_0, \|\boldsymbol{\beta}\|_0 \leq s, \|\mathbf{v}\|_1, \|\boldsymbol{\beta}\|_1 \leq L$ and $M^{-1} \leq \|\mathbf{v}\|_2, \|\boldsymbol{\beta}\|_2 \leq \delta^{-1}$. These conditions in conjunction with the assumptions of the present theorem, can be seen to imply the conditions from Theorem 4.5.9 and this completes the proof. □

Proof of Theorem 4.5.13. As in the proof of Theorem 4.5.11 it can be seen that the conditions on s and L imply the conditions of the proofs used to prove Theorem 4.5.9 hold uniformly in $\mathcal{S}_1(K, \phi, L, s)$. Thus similarly to Theorem 4.4.12 in the Dantzig Selector case, in this proof it suffices to verify that

Assumptions 4.3.21 and 4.3.18. We have

$$\begin{aligned}\sqrt{n}|S(\theta,\boldsymbol{\gamma})-S(0,\boldsymbol{\gamma})-\theta| &= \sqrt{n}\theta\left|\frac{1}{n}\mathbf{v}^T\left(\sum_{i=1}^n \boldsymbol{X}_i\boldsymbol{X}_{i,1}-\boldsymbol{\Sigma}_{\boldsymbol{X},*1}\right)\right| \\ &\leq \sqrt{n}\theta\|\mathbf{v}\|_1 \underbrace{\left\|\frac{1}{n}\sum_{i=1}^n \boldsymbol{X}_i\boldsymbol{X}_{i,1}-\boldsymbol{\Sigma}_{\boldsymbol{X},*1}\right\|_\infty}_{I} \\ &\leq \sqrt{n}KLn^{-\phi}I\end{aligned}$$

Lemma C.2.2 shows that $|I|\leq \xi\sqrt{\frac{\log d}{n}}$ for a sufficiently large $\xi>0$, and hence:

$$\lim_{n\to\infty}\inf_{\boldsymbol{\Sigma}_{\boldsymbol{X}}\in\mathcal{S}_1(K,\phi,L,s)}\mathbb{P}_{\boldsymbol{\beta}}\left(\sqrt{n}|S(\theta,\boldsymbol{\gamma})-S(0,\boldsymbol{\gamma})-\theta|\leq \xi KLn^{-\phi}\sqrt{\log d}\right)=1,$$

Next we check (4.3.14). We have:

$$\left\|\frac{1}{n}\sum_{i=1}^n \boldsymbol{X}_i\boldsymbol{X}_{i,-1}^T\boldsymbol{\gamma}-\mathbf{e}_m^T\right\|_\infty \leq \left\|\frac{1}{n}\sum_{i=1}^n \boldsymbol{X}_i^{\otimes 2}\boldsymbol{\beta}-\mathbf{e}_m^T\right\|_\infty + Kn^{-\phi}\left\|\frac{1}{n}\sum_{i=1}^n \boldsymbol{X}_i\boldsymbol{X}_{i1}\right\|_\infty.$$

We know that the first term is bounded by $\xi' L\sqrt{\frac{\log d}{n}}$ with high probability. The second term, as we already argued in the proof of Theorem 4.4.12, is $\leq M+C\sqrt{\frac{\log d}{n}}$ with high probability which concludes the proof. □

Lemma C.3.2. *Let $R_{\mathbf{v}}, R\subset\{1,\ldots,d\}$ with $|R_{\mathbf{v}}|=r_{\mathbf{v}}, |R|=r$. Then we have the following:*

$$\mathbb{E}\left\|\mathrm{Vec}[(\boldsymbol{X}^{\otimes 2}-\boldsymbol{\Sigma})_{R_{\mathbf{v}},R}]\right\|_2^3 \leq (r_{\mathbf{v}}r)^{3/2}(24K_{\boldsymbol{X}}^2)^3$$

Proof. Similarly to Lemma C.2.9 we use the following inequality:

$$\left(\frac{\sum_{k\in R_{\mathbf{v}},j\in R}(\boldsymbol{X}^k\boldsymbol{X}^j-\sigma_{kj})^2}{r_{\mathbf{v}}r}\right)^3 \leq \left(\frac{\sum_{k\in R_{\mathbf{v}},j\in R}|\boldsymbol{X}^k\boldsymbol{X}^j-\sigma_{kj}|^3}{r_{\mathbf{v}}r}\right)^2$$

This gives:

$$\sqrt{\mathbb{E}\left(\frac{\sum_{k\in R_{\mathbf{v}},j\in R}(\boldsymbol{X}^k\boldsymbol{X}^j-\sigma_{kj})^2}{r_{\mathbf{v}}r}\right)^3}\leq\sqrt{\mathbb{E}\left(\frac{\sum_{k\in R_{\mathbf{v}},j\in R}|\boldsymbol{X}^k\boldsymbol{X}^j-\sigma_{kj}|^3}{r_{\mathbf{v}}r}\right)^2}$$
$$\leq\sum_{k\in R_{\mathbf{v}},j\in R}\sqrt{\mathbb{E}\left(\frac{|\boldsymbol{X}^k\boldsymbol{X}^j-\sigma_{kj}|^6}{(r_{\mathbf{v}}r)^2}\right)}\tag{C.3.2}$$

Now recall that from (C.2.4) we have $\|\boldsymbol{X}^k\boldsymbol{X}^j\|_{\psi_1}\leq 2K_{\boldsymbol{X}}^2$, which implies that $\|\boldsymbol{X}^k\boldsymbol{X}^j-\sigma_{kj}\|_{\psi_1}\leq 4K_{\boldsymbol{X}}^2$ by the definition of ψ_1 norm (4.1.2). Then again by an application of ψ_1 definition we have:

$$\sum_{k\in R_{\mathbf{v}},j\in R}\sqrt{\mathbb{E}\left(\frac{|\boldsymbol{X}^k\boldsymbol{X}^j-\sigma_{kj}|^6}{(r_{\mathbf{v}}r)^2}\right)}\leq(24K_{\boldsymbol{X}}^2)^3\tag{C.3.3}$$

Hence since:

$$\mathbb{E}\left\|\mathrm{Vec}[(\boldsymbol{X}^{\otimes 2}-\boldsymbol{\Sigma}_{\boldsymbol{X}})_{R_{\mathbf{v}},R}]\right\|_2^3\leq\sqrt{\mathbb{E}\left(\sum_{k\in R_{\mathbf{v}},j\in R}(\boldsymbol{X}^k\boldsymbol{X}^j-\sigma_{kj})^2\right)^3},$$

we get:

$$\mathbb{E}\left\|\mathrm{Vec}[(\boldsymbol{X}^{\otimes 2}-\boldsymbol{\Sigma}_{\boldsymbol{X}})_{R_{\mathbf{v}},R}]\right\|_2^3\leq(r_{\mathbf{v}}r)^{3/2}(24K_{\boldsymbol{X}}^2)^3,$$

as claimed. □

C.3.2 Proofs for Transelliptical Models

Proof of Theorem 4.5.21. Note that by the mean value theorem we have the following representation:

$$\begin{aligned} n^{1/2}\mathbf{v}^{*T}\left(\widehat{\boldsymbol{S}}^{\tau}\boldsymbol{\beta}^{*}-\mathbf{e}_m^T\right) &= n^{1/2}\mathbf{v}^{*T}\left(\widehat{\boldsymbol{S}}^{\tau}-\boldsymbol{\Sigma}\right)\boldsymbol{\beta}^{*} \\ &= n^{1/2}\sum_{\substack{j\in S_{\mathbf{v}},k\in S\\ j\neq k}}\mathbf{v}_j^{*}\boldsymbol{\beta}_k^{*}\left(\sin\left(\widehat{\tau}_{jk}\frac{\pi}{2}\right)-\sin\left(\tau_{jk}\frac{\pi}{2}\right)\right) \\ &= n^{1/2}\sum_{\substack{j\in S_{\mathbf{v}},k\in S\\ j\neq k}}\mathbf{v}_j^{*}\boldsymbol{\beta}_k^{*}\cos\left(\tau_{jk}\frac{\pi}{2}\right)\frac{\pi}{2}\left(\widehat{\tau}_{jk}-\tau_{jk}\right) \\ &\quad -\frac{n^{1/2}}{2}\sum_{\substack{i\in S_{\mathbf{v}},j\in S\\ j\neq k}}\mathbf{v}_j^{*}\boldsymbol{\beta}_k^{*}\sin\left(\widetilde{\tau}_{jk}\frac{\pi}{2}\right)\left(\frac{\pi}{2}\left(\widehat{\tau}_{jk}-\tau_{jk}\right)\right)^2, \end{aligned}$$

where $\widetilde{\tau}_{ij}$ is a number between $\widehat{\tau}_{ij}$ and τ_{ij}. We will first deal with the first term in the sum above. Since this term is a linear combination of second order (dependent) U-statistics, we will make usage of Háejk's projection method. A similar approach was used in the celebrated paper of Hoeffding [31]. To this end we define the following notations:

$$\begin{aligned} \tau_{jk}^{ii'} &= \operatorname{sign}\left((\boldsymbol{X}_i^j-\boldsymbol{X}_{i'}^j)(\boldsymbol{X}_i^k-\boldsymbol{X}_{i'}^k)\right)-\tau_{jk}, \\ \tau_{jk}^{ii'|i} &= \mathbb{E}[\tau_{jk}^{ii'}|\boldsymbol{X}_i] \\ &= \left[\mathbb{P}\left((\boldsymbol{X}_i^j-\boldsymbol{X}^j)(\boldsymbol{X}_i^k-\boldsymbol{X}^k)>0|\boldsymbol{X}_i\right)-\mathbb{P}\left((\boldsymbol{X}_i^j-\boldsymbol{X}^j)(\boldsymbol{X}_i^k-\boldsymbol{X}^k)<0|\boldsymbol{X}_i\right)\right]-\tau_{jk}, \\ \tau_{jk}^{i} &= \frac{1}{n-1}\sum_{i'\neq i}\tau_{jk}^{ii'|i}, \\ w_{jk}^{ii'} &= \tau_{jk}^{ii'}-\tau_{jk}^{ii'|i}-\tau_{jk}^{ii'|i'}, \end{aligned}$$

where in the last line $\boldsymbol{X}_i$ is held fixed, while $\boldsymbol{X}$ is an independent copy of $\boldsymbol{X}_i$. In terms of these notations we therefore have:

$$\widehat{\tau}_{jk} - \tau_{jk} = \frac{2}{n}\sum_{i=1}^{n} \tau_{jk}^i + \frac{2}{n(n-1)} \sum_{1 \le i < i' \le n} w_{jk}^{ii'}.$$

This gives us the following identity:

$$n^{1/2} \sum_{\substack{j \in S_{\mathbf{v}}, k \in S \\ j \neq k}} \mathbf{v}_j^* \boldsymbol{\beta}_k^* \cos\left(\tau_{jk}\frac{\pi}{2}\right) \frac{\pi}{2} \left(\widehat{\tau}_{jk} - \tau_{jk}\right) = \underbrace{\pi n^{-1/2} \sum_{\substack{j \in S_{\mathbf{v}}, k \in S \\ j \neq k}} \mathbf{v}_j^* \boldsymbol{\beta}_k^* \cos\left(\tau_{jk}\frac{\pi}{2}\right) \sum_{i=1}^{n} \tau_{jk}^i}_{I_1}$$
$$+ \underbrace{\frac{\pi}{n^{1/2}(n-1)} \sum_{\substack{j \in S_{\mathbf{v}}, k \in S \\ j \neq k}} \mathbf{v}_j^* \boldsymbol{\beta}_k^* \cos\left(\tau_{jk}\frac{\pi}{2}\right) \sum_{1 \le i < i' \le n} w_{jk}^{ii'}}_{I_2}.$$

We first deal with I_1 which can clearly be represented as a sum of iid mean 0 terms, by verifying Lyapunov's condition for the CLT. I_1 can be rewritten as:

$$I_1 = n^{-1/2} \sum_{i=1}^{n} \underbrace{\sum_{\substack{j \in S_{\mathbf{v}}, k \in S \\ j \neq k}} \mathbf{v}_j^* \boldsymbol{\beta}_k^* \pi \cos\left(\tau_{jk}\frac{\pi}{2}\right) \tau_{jk}^i}_{M_i}. \tag{C.3.4}$$

Define the matrix Θ^i where the

$$\Theta_{jk}^i = \pi \cos\left(\tau_{jk}\frac{\pi}{2}\right) \tau_{jk}^i, \text{ hence } \Theta_{jj}^i = 0. \tag{C.3.5}$$

We can then rewrite $M_i = \mathbf{v}^{*T}\Theta^i\boldsymbol{\beta}^* = \mathbf{v}_{S_{\mathbf{v}}}^{*T}\Theta_{S_{\mathbf{v}},S}^i\boldsymbol{\beta}_S^*$. Calculating the variance of M_i gives:

$$\mathrm{Var}(M_i) = \mathbb{E}(\mathbf{v}^{*T}\Theta^i\boldsymbol{\beta}^*)^2 \ge \iota_{\min}\|\mathbf{v}^*\|_2^2\|\boldsymbol{\beta}^*\|_2^2$$

by our assumption. We proceed to verify Lyapunov's condition (where we ignore the constant $\iota_{\min} > 0$):

$$\frac{n^{-3/2}}{\|\mathbf{v}^*\|_2^3 \|\boldsymbol{\beta}^*\|_2^3} \sum_{i=1}^{n} \mathbb{E}|M_i|^3 \leq n^{-3/2} \sum_{i=1}^{n} \mathbb{E}\| \operatorname{Vec}(\Theta^i_{S_{\mathbf{v}},S})\|_2^3,$$

where the last inequality follows from Cauchy-Schwartz. Finally, notice that each element of Θ^i is bounded $|\Theta^i_{jk}| \leq 2\pi$, and hence $\| \operatorname{Vec}(\Theta^i_{S_{\mathbf{v}},S})\|_2^3 \leq (s_{\mathbf{v}} s)^{3/2}(2\pi)^3$. Thus finally:

$$\frac{n^{-3/2}}{\|\mathbf{v}^*\|_2^3 \|\boldsymbol{\beta}^*\|_2^3} \sum_{i=1}^{n} \mathbb{E}|M_i|^3 \leq \frac{(s_{\mathbf{v}} s)^{3/2}(2\pi)^3}{n^{1/2}} = o(1)$$

The last equality follows from our assumption. This implies that $I_1 \rightsquigarrow N(0, \Delta)$, with $\Delta = \mathbb{E}(\mathbf{v}^{*T}\Theta^i\boldsymbol{\beta}^*)^2$.

Next we deal with the second term I_2, which is also unbiased, by showing that its (standardized) variance goes to 0 asymptotically. Before we compute its variance we make several preliminary calculations:

$$\begin{aligned}
\mathbb{E}(w^{ii'}_{jk} w^{rr'}_{ls}) = \mathbb{E}(\tau^{ii'}_{jk}\tau^{rr'}_{ls}) &- \mathbb{E}(\tau^{ii'}_{jk}\tau^{rr'|r}_{ls}) - \mathbb{E}(\tau^{ii'}_{jk}\tau^{rr'|r'}_{ls}) \\
&- \mathbb{E}(\tau^{ii'|i}_{jk}\tau^{rr'}_{ls}) + \mathbb{E}(\tau^{ii'|i}_{jk}\tau^{rr'|r}_{ls}) + \mathbb{E}(\tau^{ii'|i}_{jk}\tau^{rr'|r'}_{ls}) \\
&- \mathbb{E}(\tau^{ii'|i'}_{jk}\tau^{rr'}_{ls}) + \mathbb{E}(\tau^{ii'|i'}_{jk}\tau^{rr'|r}_{ls}) + \mathbb{E}(\tau^{ii'|i'}_{jk}\tau^{rr'|r'}_{ls}).
\end{aligned}$$

In the expression above we have taken $j \neq k, l \neq s, r \neq i \neq i' \neq r \neq r' \neq i$. Notice now that all elements above are independent and since $\mathbb{E}(w^{ii'}_{jk}) = \mathbb{E}(w^{rr'}_{ls}) = 0$, we conclude that $\mathbb{E}(w^{ii'}_{jk} w^{rr'}_{ls}) = 0$. Following, the same logic, for $j \neq k, l \neq s, i \neq i' \neq r' \neq i$:

$$\mathbb{E}(w^{ii'}_{jk} w^{ir'}_{ls}) = \mathbb{E}(\tau^{ii'}_{jk}\tau^{ir'}_{ls}) - \mathbb{E}(\tau^{ii'}_{jk}\tau^{ir'|i}_{ls}) - \mathbb{E}(\tau^{ii'|i}_{jk}\tau^{ir'}_{ls}) + \mathbb{E}(\tau^{ii'|i}_{jk}\tau^{ir'|i}_{ls})$$

where all the rest terms are 0, by the same argument as in the first case. Using iterated expectation by

conditioning on $\boldsymbol{X}_i$ it can be easily seen that all terms become equal to — $\mathbb{E}(\tau_{jk}^{ii'|i}\tau_{jk}^{ir'|i})$, and we can conclude that $\mathbb{E}(w_{jk}^{ii'}w_{ls}^{ir'}) = 0$.

Since $\mathbb{E}I_2 = 0$, we have:

$$\begin{aligned}\frac{\mathrm{Var}(I_2)}{\mathrm{Var}(M_i)} &\leq \frac{\mathbb{E}(I_2^2)}{\iota_{\min}\|\mathbf{v}^*\|_2^2\|\boldsymbol{\beta}^*\|_2^2} = \frac{\pi^2}{\iota_{\min}\|\mathbf{v}^*\|_2^2\|\boldsymbol{\beta}^*\|_2^2 n(n-1)^2}\sum_{1\leq i<i'\leq n}\mathbb{E}\left(\sum_{\substack{j\in S_{\mathbf{v}},k\in S\\ j\neq k}}\mathbf{v}_j^*\boldsymbol{\beta}_k^*\cos\left(\tau_{jk}\frac{\pi}{2}\right)w_{jk}^{ii'}\right)^2 \\ &\leq \frac{\pi^2\binom{n}{2}36\left(\sum_{j\in S_{\mathbf{v}}}|\mathbf{v}_j^*|\right)^2\left(\sum_{k\in S}|\boldsymbol{\beta}_k^*|\right)^2}{n(n-1)^2\iota_{\min}\|\mathbf{v}^*\|_2^2\|\boldsymbol{\beta}^*\|_2^2} \\ &\leq \frac{\pi^2 18 s_{\mathbf{v}}s}{(n-1)\iota_{\min}} = o(1),\end{aligned}$$

where in the next to last inequality we used the trivial bound $|w_{jk}^{ii'}| \leq |\tau_{jk}^{ii'}| + |\tau_{jk}^{ii'|i}| + |\tau_{jk}^{ii'|i'}| \leq 6$. Thus the term $\frac{\mathrm{Var}(I_2)}{\mathrm{Var}(M_i)} = o(1)$ and therefore, Chebyshev's inequality gives us that $\frac{I_2}{\sqrt{\mathrm{Var}(M_i)}} = o_p(1)$.

Finally we deal with the standardized version of the last term:

$$\frac{1}{\sqrt{\mathrm{Var}(\mathbf{v}^{*T}\Theta\boldsymbol{\beta}^*)}}\frac{n^{1/2}}{2}\sum_{\substack{i\in S_{\mathbf{v}},j\in S\\ j\neq k}}\mathbf{v}_j^*\boldsymbol{\beta}_k^*\sin\left(\widetilde{\tau}_{jk}\frac{\pi}{2}\right)\left(\frac{\pi}{2}\left(\widehat{\tau}_{jk}-\tau_{jk}\right)\right)^2. \tag{C.3.6}$$

As we mentioned previously it's clear that $\widehat{\tau}_{jk}$ is a U-statistic, and its kernel is a bounded function (between -1 and 1). Furthermore, we have that $\mathbb{E}\widehat{\tau}_{jk} = \tau_{jk}$. Thus, we can apply Hoeffding's inequality for U-statistics (see Hoeffding[32] equation (5.7)), to obtain that:

$$\mathbb{P}(\sup_{jk}|\widehat{\tau}_{jk}-\tau_{jk}| > t) \leq 2d^2\exp\left(-\frac{nt^2}{4}\right). \tag{C.3.7}$$

It follows that selecting $t = 9\sqrt{\frac{\log d}{n}}$ suffices to keep the probability going to 0. Notice that the

(C.3.6) can be controlled by:

$$\frac{n^{1/2}\pi^2\sqrt{s_{\mathbf{v}}s}}{8\theta_{\min}\|\mathbf{v}^*\|_2\|\boldsymbol{\beta}^*\|_2}\|\mathbf{v}^*\|_2\|\boldsymbol{\beta}^*\|_2\sup_{jk}(\widehat{\tau}_{jk}-\tau_{jk})^2 = O_p\left(\frac{\sqrt{s_{\mathbf{v}}s}\log d}{n^{1/2}}\right) = o_p(1).$$

This concludes the proof. □

Remark C.3.3. *Using the Berry-Esseen theorem for non-identical random variables we can strengthen weak convergence statement to:*

$$\sup_t\left|\mathbb{P}^*\left(\frac{I_1}{\sqrt{\Delta}}\leq t\right)-\Phi(t)\right| \leq C_{BE}n^{-1/2}(s_{\mathbf{v}}s)^{3/2} = o(1).$$

where C_{BE} is an absolute constant. Note that we decomposed our test into $\frac{I_1}{\sqrt{\Delta}}+o_p(1)$, and hence this statement is valid more generally for Theorem 4.5.21.

Proof of Proposition 4.5.23. Before we go to the main proof, recall the definition of Θ^i (C.3.5), where $\Theta^i_{jk} = \pi\cos\left(\tau_{jk}\frac{\pi}{2}\right)\tau^i_{jk}$. Note that in fact $\mathbb{E}\Theta^i = 0$, since $\mathbb{E}\tau^i_{jk} = 0$, and thus $\mathrm{Var}(\mathbf{v}^{*T}\Theta^i\boldsymbol{\beta}^*) = \mathbb{E}(\mathbf{v}^{*T}\Theta^i\boldsymbol{\beta}^*)^2$. Similarly one can note the simple identity: $\frac{1}{n}\sum_{i=1}^n\widehat{\Theta}^i = 0$. Thus we will in fact focus on showing that $\frac{1}{n}\sum_{i=1}^n(\widehat{\mathbf{v}}^T\widehat{\Theta}^i\widehat{\boldsymbol{\beta}})^2$ is consistent for $\mathbb{E}(\mathbf{v}^{*T}\Theta^i\boldsymbol{\beta}^*)^2$.

Consider the following decomposition:

$$\frac{1}{n}\sum_{i=1}^n(\widehat{\mathbf{v}}^T\widehat{\Theta}^i\widehat{\boldsymbol{\beta}})^2 = \underbrace{\frac{1}{n}\sum_{i=1}^n[(\widehat{\mathbf{v}}^T\widehat{\Theta}^i\widehat{\boldsymbol{\beta}})^2-(\mathbf{v}^{*T}\widehat{\Theta}^i\widehat{\boldsymbol{\beta}})^2]}_{I_1} + \underbrace{\frac{1}{n}\sum_{i=1}^n[(\mathbf{v}^{*T}\widehat{\Theta}^i\widehat{\boldsymbol{\beta}})^2-(\mathbf{v}^{*T}\widehat{\Theta}^i\boldsymbol{\beta}^*)^2]}_{I_2}.$$

Below we show that I_1 is asymptotically negligible.

$$I_1 = (\widehat{\mathbf{v}}-\mathbf{v}^*)^T\frac{1}{n}\sum_{i=1}^n\widehat{\Theta}^i\widehat{\boldsymbol{\beta}}\widehat{\boldsymbol{\beta}}^T\widehat{\Theta}^i(\widehat{\mathbf{v}}+\mathbf{v}^*)^T.$$

Note that $\|\widehat{\Theta}^i\|_{\max} \leq 2\pi$, and thus:

$$|I_1| \leq \|\widehat{\mathbf{v}} - \mathbf{v}^*\|_1 \|\widehat{\mathbf{v}} + \mathbf{v}^*\|_1 \|\widehat{\boldsymbol{\beta}}\|_1^2 (2\pi)^2.$$

Using the help of Lemma C.3.6, we can get the following:

$$|I_1| = O_p\left(\|\mathbf{v}^*\|_1^2 \|\boldsymbol{\beta}^*\|_1^2 s_{\mathbf{v}} \sqrt{\frac{\log d}{n}}\right) = o_p(1),$$

by assumption. Similarly we get:

$$|I_2| = O_p\left(\|\mathbf{v}^*\|_1^2 \|\boldsymbol{\beta}^*\|_1^2 s \sqrt{\frac{\log d}{n}}\right) = o_p(1).$$

Next, we inspect the following difference:

$$I_3 = \frac{1}{n}\sum_{i=1}^{n} [(\mathbf{v}^{*T}\widehat{\Theta}^i\boldsymbol{\beta}^*)^2 - (\mathbf{v}^{*T}\Theta^i\boldsymbol{\beta}^*)^2].$$

Before we bound this term recall that we have the following useful inequality $\|\Theta^i\|_{\max} \leq 2\pi$.

Thus:

$$|I_3| \leq \|\mathbf{v}^*\|_1^2 \|\boldsymbol{\beta}^*\|_1^2 4\pi \max_{i=1,\ldots,n} \|\widehat{\Theta}^i - \Theta^i\|_{\max}. \tag{C.3.8}$$

To bound the difference $\max_{i=1,\ldots,n} \|\widehat{\Theta}^i - \Theta^i\|_{\max}$ we will use some concentration inequalities. First, since cos is Lipchitz with constant 1 — $\left|\cos\left(\frac{\pi}{2}\widehat{\tau}_{jk}\right) - \cos\left(\frac{\pi}{2}\tau_{jk}\right)\right| \leq \frac{\pi}{2}|\widehat{\tau}_{jk} - \tau_{jk}|$, we have:

$$\begin{aligned}
|\widehat{\Theta}^i_{jk} - \Theta^i_{jk}| &\leq \pi \left|\cos\left(\frac{\pi}{2}\widehat{\tau}_{jk}\right) - \cos\left(\frac{\pi}{2}\tau_{jk}\right)\right| |\widehat{\tau}^i_{jk}| + \pi \left|\cos\left(\frac{\pi}{2}\tau_{jk}\right)\right| |\widehat{\tau}^i_{jk} - \tau^i_{jk}| \\
&\leq \pi^2 |\widehat{\tau}_{jk} - \tau_{jk}| + \pi |\widehat{\tau}^i_{jk} - \tau^i_{jk}|.
\end{aligned}$$

where we used the simple observation that $|\widehat{\tau}^i_{jk}| \leq 2$. Next we have:

$$|\widehat{\tau}^i_{jk} - \tau^i_{jk}| \leq |\widehat{\tau}_{jk} - \tau_{jk}| + \Bigg| \underbrace{\frac{1}{n-1}\sum_{i' \neq i} \operatorname{sign}\left((\boldsymbol{X}^j_i - \boldsymbol{X}^j_{i'})(\boldsymbol{X}^k_i - \boldsymbol{X}^k_{i'})\right)}_{\widehat{\theta}^i_{jk}} - \underbrace{\mathbb{E}\left[\operatorname{sign}\left((\boldsymbol{X}^j_i - \boldsymbol{X}^j_{i'})(\boldsymbol{X}^k_i - \boldsymbol{X}^k_{i'})\right) | \boldsymbol{X}_i\right]}_{\theta^i_{jk}} \Bigg|$$

This gives us:

$$|\widehat{\Theta}^i_{jk} - \Theta^i_{jk}| \leq (\pi^2 + \pi) \sup_{jk} |\widehat{\tau}_{jk} - \tau_{jk}| + \pi |\widehat{\theta}^i_{jk} - \theta^i_{jk}|. \tag{C.3.9}$$

Next, note since the terms in $\widehat{\theta}^i_{jk}$ are iid conditional on $\boldsymbol{X}_i$, and they are in the set $\{-1, 1\}$ by Hoeffding's inequality:

$$\mathbb{P}(|\widehat{\theta}^i_{jk} - \theta^i_{jk}| > t | \boldsymbol{X}_i) \leq 2 \exp\left(-\frac{(n-1)t^2}{2}\right).$$

Of course, it follows that the same inequality holds unconditionally as well, so:

$$\mathbb{P}(|\widehat{\theta}^i_{jk} - \theta^i_{jk}| > t) \leq 2 \exp\left(-\frac{(n-1)t^2}{2}\right).$$

Applying the union bound over all i, j, k we get that:

$$\mathbb{P}(\max_{i,j,k} |\widehat{\theta}^i_{jk} - \theta^i_{jk}| > t) \leq 2nd^2 \exp\left(-\frac{(n-1)t^2}{2}\right).$$

This implies that selecting $t = 4\sqrt{\frac{\log(nd)}{n}}$, would keep the probability converging to 0. Recall that by (C.3.7) we have already observed that:

$$\mathbb{P}(\sup_{jk} |\widehat{\tau}_{jk} - \tau_{jk}| > t) \leq 2d^2 \exp\left(-\frac{nt^2}{4}\right).$$

and similarly as we observed before we can set $t = 9\sqrt{\frac{\log d}{n}}$, to keep the tail bound converging to 0. To summarize the above inequalities and (C.3.9) give us:

$$\max_i |\widehat{\Theta}^i_{jk} - \Theta^i_{jk}|_{\max} = O_p\left(\sqrt{\frac{\log(nd)}{n}}\right).$$

Thus using (C.3.8), we get:

$$|I_3| = \|\mathbf{v}^*\|_1^2 \|\boldsymbol{\beta}^*\|_1^2 O_p\left(\sqrt{\frac{\log(nd)}{n}}\right) = o_p(1).$$

Finally we asses the difference:

$$\frac{1}{n}\sum_{i=1}^{n}(\mathbf{v}^{*T}\Theta^i\boldsymbol{\beta}^*)^2 - \mathbb{E}(\mathbf{v}^{*T}\Theta^i\boldsymbol{\beta}^*)^2.$$

By Chebyshev's inequality in much the same way as in the last part of the proof of Proposition 4.5.6, we can show that the expression above is $o_p(1)$ under the assumption $\mathrm{Var}((\mathbf{v}^{*T}\Theta\boldsymbol{\beta}^*)^2) = o(n)$.

□

Proof of Theorem 4.5.26. Similarly to the proof of Theorem 4.5.11 we simply need to note that our conditions imply the conditions required by Theorem 4.5.25 and also note that the bounds in the proofs hold uniformly. □

Proof of Theorem 4.5.27. In this proof we show that the uniform local approximation (Assumption 4.3.21) holds. It is clear that:

$$\begin{aligned}
\sqrt{n}|S(\theta,\boldsymbol{\gamma}) - S(0,\boldsymbol{\gamma}) - \theta| &= \sqrt{n}|\theta\mathbf{v}^T(\widehat{S}^{\tau}_{*1} - \boldsymbol{\Sigma}_{*1})| \\
&\leq \sqrt{n}Kn^{-\phi}\|\mathbf{v}\|_1\|\widehat{S}^{\tau}_{*1} - \boldsymbol{\Sigma}_{*1}\|_\infty \\
&\leq \sqrt{n}Kn^{-\phi}L\|\widehat{S}^{\tau} - \boldsymbol{\Sigma}\|_{\max}.
\end{aligned}$$

Now from Theorem 4.5.19 we can conclude that with high probability the last expression is bounded by $2.45\pi K n^{-\phi} L\sqrt{\log d} \to 0$ by assumption.

Next we show (4.3.9). We have:

$$\left\|\widehat{\boldsymbol{S}}^{\tau}(0,\boldsymbol{\gamma}^T)^T - \mathbf{e}_m^T\right\|_\infty \leq \left\|\widehat{\boldsymbol{S}}^{\tau}\boldsymbol{\beta} - \mathbf{e}_m^T\right\|_\infty + Kn^{-\phi}\left\|\widehat{\boldsymbol{S}}^{\tau}_{*1}\right\|_\infty.$$

We know that the first term is bounded by $\xi' L\sqrt{\frac{\log d}{n}}$ with high probability, and for the second term we can argue similarly to the proof of Theorem 4.4.12, is $\leq 1 + C\sqrt{\frac{\log d}{n}}$ with high probability which concludes the proof. □

Lemma C.3.4. *Assume that the minimum eigenvalue* $\lambda_{\min}(\boldsymbol{\Sigma}) > 0$ *and* $s\sqrt{\frac{\log d}{n}} \leq (1-\kappa)\frac{\lambda_{\min}(\boldsymbol{\Sigma}_{\boldsymbol{X}})}{(1+\xi)^2 2.45\pi}$*, where* $0 < \kappa < 1$*. We then have that* $\widehat{\boldsymbol{S}}^{\tau}$ *satisfies the RE property with* $\mathrm{RE}_{\widehat{\boldsymbol{S}}^{\tau}}(s,\xi) \geq \kappa\lambda_{\min}(\boldsymbol{\Sigma})$ *with probability at least* $1 - 1/d$*.*

Proof of Lemma C.3.4. Proof is the same as in Lemma C.2.3, but we use Theorem 4.5.19 instead of Lemma C.2.2. Thus we omit it. □

Definition C.3.5. *Define* $\mathrm{RE}_\kappa(s,\xi) := \kappa\,\mathrm{RE}_{\boldsymbol{\Sigma}}(s,\xi) \geq \kappa\lambda_{\min}(\boldsymbol{\Sigma})$.

Lemma C.3.6. *Assume that —* $\lambda_{\min}(\boldsymbol{\Sigma}) > 0$*,* $s_{\mathbf{v}}\sqrt{\frac{\log d}{n}} \leq (1-\kappa)\frac{\lambda_{\min}(\boldsymbol{\Sigma})}{(1+1)^2 2.45\pi}$*, where* $0 < \kappa < 1$ *and* $\lambda' \geq \|\mathbf{v}^*\|_1 2.45\pi\sqrt{\frac{\log d}{n}}$*. Then we have that* $\|\widehat{\mathbf{v}} - \mathbf{v}^*\|_1 \leq \frac{8\lambda' s_{\mathbf{v}}}{\mathrm{RE}_\kappa(s_{\mathbf{v}},1)}$ *with probability at least* $1 - 1/d$*.*

Proof of Lemma C.3.6. Proof is the same as in Lemma C.2.5, but we use Theorem 4.5.19 instead of Lemma C.2.2 and we use Lemma C.3.4 instead of Lemma C.2.3. Thus we omit it. □

C.4 Proofs for the LDP Inference

Proof of Theorem 4.6.1. Let $\widehat{\boldsymbol{\beta}}_0 = (0, \widehat{\boldsymbol{\gamma}}^T)^T$. We have the following identity:

$$n^{1/2}\widehat{\mathbf{v}}^T(\widehat{\boldsymbol{\Sigma}}_n\widehat{\boldsymbol{\beta}}_0 - (\bar{\boldsymbol{X}} - \bar{\boldsymbol{Y}})) = \underbrace{n^{1/2}\widehat{\mathbf{v}}^T(\widehat{\boldsymbol{\Sigma}}_n\boldsymbol{\beta}^* - (\bar{\boldsymbol{X}} - \bar{\boldsymbol{Y}}))}_{I_1} + \underbrace{n^{1/2}\widehat{\mathbf{v}}^T\widehat{\boldsymbol{\Sigma}}_n(\widehat{\boldsymbol{\beta}}_0 - \boldsymbol{\beta}^*)}_{I_2}.$$

Before we proceed with expanding the first term, let us define the following quantity:

$$\widetilde{\boldsymbol{\Sigma}}_n = \frac{1}{n}\left[\sum_{i=1}^{n_1}(\boldsymbol{X}_i - \boldsymbol{\mu}_1)^{\otimes 2} + \sum_{i=1}^{n_2}(\boldsymbol{Y}_i - \boldsymbol{\mu}_2)^{\otimes 2}\right]. \tag{C.4.1}$$

We then have the following identity:

$$I_1 = \underbrace{n^{1/2}\mathbf{v}^{*T}(\widetilde{\boldsymbol{\Sigma}}_n\boldsymbol{\beta}^* - (\bar{\boldsymbol{X}} - \bar{\boldsymbol{Y}}))}_{I_{11}} + n^{1/2}\underbrace{\mathbf{v}^{*T}(\widehat{\boldsymbol{\Sigma}}_n - \widetilde{\boldsymbol{\Sigma}}_n)\boldsymbol{\beta}^*}_{I_{12}} + \underbrace{n^{1/2}(\widehat{\mathbf{v}} - \mathbf{v}^*)^T(\widehat{\boldsymbol{\Sigma}}_n\boldsymbol{\beta}^* - (\bar{\boldsymbol{X}} - \bar{\boldsymbol{Y}}))}_{I_{13}}.$$

We proceed with the terms I_{12} and I_{13}, showing that both terms are small.

$$|I_{12}| \leq n^{1/2}\|\mathbf{v}^*\|_1\|\boldsymbol{\beta}^*\|_1\|\widehat{\boldsymbol{\Sigma}}_n - \widetilde{\boldsymbol{\Sigma}}_n\|_{\max} = \|\mathbf{v}^*\|_1\|\boldsymbol{\beta}^*\|_1 O_p\left(\frac{\log d}{n^{1/2}}\right) = o_p(1),$$

where we used (C.4.3) from Lemma C.4.1 (and made usage of the fact that $n_1 \asymp n_2$). We can control I_{13} by the following:

$$\begin{aligned}
|I_{13}| &\leq n^{1/2}\|\widehat{\mathbf{v}} - \mathbf{v}^*\|_1\|\widehat{\boldsymbol{\Sigma}}_n\boldsymbol{\beta}^* - (\bar{\boldsymbol{X}} - \bar{\boldsymbol{Y}}))\|_\infty \\
&\leq n^{1/2}\|\mathbf{v}^*\|_1 O_p\left(s_{\mathbf{v}}\sqrt{\frac{\log d}{n}}\right)(\|\boldsymbol{\beta}^*\|_1 \vee 1)O_p\left(\sqrt{\frac{\log d}{n}}\right) \\
&= o_p(1),
\end{aligned}$$

where the next to last inequality follows from Lemma C.4.4 and Lemma C.4.6. We next deal with I_2:

$$\begin{aligned}
|I_2| &\leq n^{1/2}\|\widehat{\mathbf{v}}^T(\widehat{\boldsymbol{\Sigma}}_n)_{-1}\|_\infty\|\widehat{\boldsymbol{\beta}}_0 - \boldsymbol{\beta}^*\|_1 \\
&\leq n^{1/2}\lambda'\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_1 \\
&\leq n^{1/2}\|\mathbf{v}^*\|_1 O_p\left(\sqrt{\frac{\log d}{n}}\right)(\|\boldsymbol{\beta}^*\|_1 \vee 1)O_p\left(s\sqrt{\frac{\log d}{n}}\right) \\
&= o_p(1),
\end{aligned}$$

where $(\widehat{\boldsymbol{\Sigma}}_n)_{-1}$ means dropping the first column (the one corresponding to the 0 coefficient in $\boldsymbol{\beta}^*$ under the null) of $\widehat{\boldsymbol{\Sigma}}_n$, as by definition $\widehat{\boldsymbol{\beta}}_0 = (0, \widehat{\boldsymbol{\gamma}}^T)^T$ and $\boldsymbol{\beta}^* = (0, \boldsymbol{\gamma}^{*T})^T$ under the null. The second inequality in the preceding display follows from Lemma C.4.4 and Lemma C.4.6. Next we take a closer look at the term I_{11}:

$$\begin{aligned}
I_{11} &= n^{1/2}\mathbf{v}^{*T}(\widetilde{\boldsymbol{\Sigma}}_n\boldsymbol{\beta}^* - (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)) + n^{1/2}\mathbf{v}^{*T}(\bar{\boldsymbol{X}} - \boldsymbol{\mu}_1 - \bar{\boldsymbol{Y}} + \boldsymbol{\mu}_2) \\
&= n^{1/2}\mathbf{v}^{*T}\frac{1}{n}\sum_{i=1}^{n}\left(\boldsymbol{U}_i^{\otimes 2}\boldsymbol{\beta}^* - (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) + \left[\frac{n}{n_1}I(i \leq n_1) - \frac{n}{n_2}I(i > n_1)\right]\boldsymbol{U}_i\right).
\end{aligned}$$

This completes the proof. □

Proof of Corollary 4.6.3. First, from Theorem 4.6.1 and $n_1 - n\alpha = o(1)$, it is clear that:

$$\begin{aligned}
n^{1/2}\widehat{S}(0, \widehat{\boldsymbol{\gamma}}) = &\frac{1}{n^{1/2}}\mathbf{v}^{*T}\sum_{i=1}^{n_1}\left(\boldsymbol{U}_i^{\otimes 2}\boldsymbol{\beta}^* - (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) + \alpha^{-1}\boldsymbol{U}_i\right) \\
&+ \frac{1}{n^{1/2}}\mathbf{v}^{*T}\sum_{i=n_1+1}^{n}\left(\boldsymbol{U}_i^{\otimes 2}\boldsymbol{\beta}^* - (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) - (1-\alpha)^{-1}\boldsymbol{U}_i\right) + o_p(1),
\end{aligned}$$

where implicitly used Chebyshev's inequality and the fact that $\mathrm{Var}(\mathbf{v}^{*T}\boldsymbol{U}) \leq 2\mathbf{v}^{*T}\boldsymbol{\Sigma}\mathbf{v}^* \leq 2\delta^{-1}$.

Next we verify Lyapunov's condition. The sum of variances of the terms above equals:

$$n_1 V_1 + n_2 V_2 = n(\alpha V_1 + (1-\alpha)V_2)(1+o(1)) \geq n V_{\min}(\|\boldsymbol{\beta}^*\|_2^2 \|\mathbf{v}^*\|_2^2 + \|\mathbf{v}^*\|_2^2)(1+o(1)),$$

by (4.6.2). Without loss of generality let's assume that $\alpha^{-1} > (1-\alpha)^{-1}$. It follows then from Lemma C.4.8, that:

$$\mathbb{E}\left|\mathbf{v}^{*T}\boldsymbol{U}_i^{\otimes 2}\boldsymbol{\beta}^* - \mathbf{v}^{*T}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) + \alpha^{-1}\mathbf{v}^{*T}\boldsymbol{U}_i\right|^3 \leq \|\mathbf{v}^*\|_2^3 (C_1 (s_{\mathbf{v}} s)^{3/2} \|\boldsymbol{\beta}^*\|_2^3 + C_2 \alpha^{-3} s_{\mathbf{v}}^{3/2}),$$

and similarly:

$$\mathbb{E}\left|\mathbf{v}^{*T}\boldsymbol{U}_i^{\otimes 2}\boldsymbol{\beta}^* - \mathbf{v}^{*T}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) - (1-\alpha)^{-1}\mathbf{v}^{*T}\boldsymbol{U}_i\right|^3 \leq \|\mathbf{v}^*\|_2^3 (C_1 (s_{\mathbf{v}} s)^{3/2} \|\boldsymbol{\beta}^*\|_2^3 + C_2 \alpha^{-3} s_{\mathbf{v}}^{3/2}),$$

where C_1 and C_2 are some absolute constants (see the Lemma for details). Therefore we conclude that the sum in Lyapunov's condition, is bounded by:

$$\frac{(s_{\mathbf{v}} s)^{3/2}}{(1+o(1))n^{1/2}} \underbrace{\frac{C_1 \|\boldsymbol{\beta}^*\|_2^3 + \frac{C_2 \alpha^{-3}}{s^{3/2}}}{V_{\min}^{3/2}(\|\boldsymbol{\beta}^*\|_2^2 + 1)^{3/2}}}_{O(1)} = o(1).$$

This completes the proof. □

Proof of Proposition 4.6.5. First note that Δ can be correspondingly decomposed to $\widehat{\Delta}$ as :

$$\begin{aligned}\Delta = {} & \alpha \mathbb{E}(\mathbf{v}^{*T}\boldsymbol{U}^{\otimes 2}\boldsymbol{\beta}^*)^2 + \alpha^{-1}\mathbb{E}(\mathbf{v}^{*T}\boldsymbol{U})^2 \\ & + (1-\alpha)\mathbb{E}(\mathbf{v}^{*T}\boldsymbol{U}^{\otimes 2}\boldsymbol{\beta}^*)^2 + (1-\alpha)^{-1}\mathbb{E}(\mathbf{v}^{*T}\boldsymbol{U})^2 - (\mathbf{v}^{*T}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2))^2.\end{aligned}$$

We start from the last term:

$$\underbrace{(\widehat{\mathbf{v}}^T(\bar{\boldsymbol{X}}-\bar{\boldsymbol{Y}}))^2}_{I} = \underbrace{[(\widehat{\mathbf{v}}^T(\bar{\boldsymbol{X}}-\bar{\boldsymbol{Y}}))^2-(\mathbf{v}^{*T}(\bar{\boldsymbol{X}}-\bar{\boldsymbol{Y}}))^2]}_{I_1}+\underbrace{(\mathbf{v}^{*T}(\bar{\boldsymbol{X}}-\bar{\boldsymbol{Y}}))^2}_{I_2}.$$

We have:

$$|I_1| \leq \|\widehat{\mathbf{v}}-\mathbf{v}^*\|_1\|\widehat{\mathbf{v}}+\mathbf{v}^*\|_1\|(\bar{\boldsymbol{X}}-\bar{\boldsymbol{Y}})^{\otimes 2}\|_{\max}.$$

Using Lemma C.4.6, we know $\|\widehat{\mathbf{v}}-\mathbf{v}^*\|_1 = O_p\left(\|\mathbf{v}^*\|_1 s_{\mathbf{v}}\sqrt{\frac{\log d}{n}}\right)$. We can apply the concentration inequality (C.4.2) provided in Lemma C.4.1 to claim that:

$$\|(\bar{\boldsymbol{X}}-\bar{\boldsymbol{Y}})^{\otimes 2}\|_{\max} \leq \|\boldsymbol{\mu}_1-\boldsymbol{\mu}_2\|_\infty^2 + \|\boldsymbol{\mu}_1-\boldsymbol{\mu}_2\|_\infty O_p\left(\sqrt{\frac{\log d}{n}}\right),$$

where we used the triangle inequality $\|\bar{\boldsymbol{X}}-\bar{\boldsymbol{Y}}\|_\infty \leq \|\bar{\boldsymbol{X}}-\boldsymbol{\mu}_1\|_\infty + \|\bar{\boldsymbol{Y}}-\boldsymbol{\mu}_2\|_\infty + \|\boldsymbol{\mu}_1-\boldsymbol{\mu}_2\|_\infty$. Finally due to our assumptions we have:

$$|I_1| = \|\boldsymbol{\mu}_1-\boldsymbol{\mu}_2\|_\infty^2 O_p\left(\|\mathbf{v}^*\|_1^2 s_{\mathbf{v}}\sqrt{\frac{\log d}{n}}\right) = o_p(1).$$

Next we tackle I_2:

$$I_2 = \underbrace{(\mathbf{v}^{*T}(\bar{\boldsymbol{X}}-\bar{\boldsymbol{Y}}))^2-(\mathbf{v}^{*T}(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2))^2}_{I_{21}}+\underbrace{(\mathbf{v}^{*T}(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2))^2}_{I_{22}}.$$

In a similar fashion as before, applying the concentration inequality (C.4.2), we can get:

$$|I_{21}| \leq \|\mathbf{v}^*\|_1^2 O_p\left(\sqrt{\frac{\log d}{n}}\right)\|\boldsymbol{\mu}_1-\boldsymbol{\mu}_2\|_\infty = o_p(1),$$

by assumption. Thus we have shown:

$$I = (\mathbf{v}^{*T}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2))^2 + o_p(1).$$

To this end define the following shorthand notations:

$$I_X(\mathbf{v}, \boldsymbol{\beta}) = \frac{1}{n}\sum_{i=1}^{n_1}\left(\mathbf{v}^T(\boldsymbol{X}_i - \bar{\boldsymbol{X}})^{\otimes 2}\boldsymbol{\beta}\right)^2, \quad I_Y(\mathbf{v}, \boldsymbol{\beta}) = \frac{1}{n}\sum_{i=n_1+1}^{n}\left(\mathbf{v}^T(\boldsymbol{Y}_i - \bar{\boldsymbol{Y}})^{\otimes 2}\boldsymbol{\beta}\right)^2$$

Next we show that $I_X + I_Y$ is consistent for $\mathbb{E}(\mathbf{v}^*\boldsymbol{U}^{\otimes 2}\boldsymbol{\beta}^*)^2$. We first consider the difference:

$$\begin{aligned}
&|I_X(\widehat{\mathbf{v}}, \widehat{\boldsymbol{\beta}}) + I_Y(\widehat{\mathbf{v}}, \widehat{\boldsymbol{\beta}}) - I_X(\mathbf{v}^*, \widehat{\boldsymbol{\beta}}) - I_Y(\mathbf{v}^*, \widehat{\boldsymbol{\beta}})| \\
&\qquad \leq \|\widehat{\mathbf{v}} - \mathbf{v}^*\|_1\|\widehat{\mathbf{v}} + \mathbf{v}^*\|_1 M\|\widehat{\boldsymbol{\Sigma}}_n\widehat{\boldsymbol{\beta}}\|_\infty\|\widehat{\boldsymbol{\beta}}\|_1,
\end{aligned}$$

where:

$$M = \max\left\{\max_{i=1,\ldots,n_1}\|(\boldsymbol{X}_i - \bar{\boldsymbol{X}})^{\otimes 2}\|_{\max}, \max_{i=n_1+1,\ldots,n}\|(\boldsymbol{Y}_i - \bar{\boldsymbol{Y}})^{\otimes 2}\|_{\max}\right\}.$$

Note now that the random variables $\boldsymbol{X}_i - \bar{\boldsymbol{X}}$ and $\boldsymbol{Y}_i - \bar{\boldsymbol{Y}}$ are in fact mean 0 sub-Gaussian variables since e.g. $\|\boldsymbol{X}_i - \bar{\boldsymbol{X}}\|_{\psi_2} \leq \|\boldsymbol{X}_i - \boldsymbol{\mu}_1\|_{\psi_2} + \|\bar{\boldsymbol{X}} - \boldsymbol{\mu}_1\|_{\psi_2} \leq 2K_{\boldsymbol{U}}$. Thus an application of Lemma C.2.8, and the fact that $n_1 \asymp n_2 \asymp n$, gives us that $M = O(\log(nd))$. Furthermore we have:

$$\|\widehat{\boldsymbol{\Sigma}}_n\widehat{\boldsymbol{\beta}}\|_\infty \leq \lambda + \|\bar{\boldsymbol{X}} - \boldsymbol{\mu}_1\|_\infty + \|\bar{\boldsymbol{Y}} - \boldsymbol{\mu}_2\|_\infty + \|\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2\|_\infty.$$

An application of (C.4.2), and the way we select λ we have:

$$\|\widehat{\boldsymbol{\Sigma}}_n\widehat{\boldsymbol{\beta}}\|_\infty \leq (\|\boldsymbol{\beta}^*\|_1 \vee 1)O_p\left(\sqrt{\frac{\log d}{n}}\right) + \|\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2\|_\infty.$$

Putting the last several inequalities together with Lemma C.4.4 and Lemma C.4.6, we have:

$$\begin{aligned}
&|I_X(\widehat{\mathbf{v}}, \widehat{\boldsymbol{\beta}}) + I_Y(\widehat{\mathbf{v}}, \widehat{\boldsymbol{\beta}}) - I_X(\mathbf{v}^*, \widehat{\boldsymbol{\beta}}) - I_Y(\mathbf{v}^*, \widehat{\boldsymbol{\beta}})| \\
&\leq \|\mathbf{v}^*\|_1^2 \|\boldsymbol{\beta}^*\|_1 s_{\mathbf{v}} \log(nd) O_p\left(\sqrt{\frac{\log d}{n}}\right)\left[(\|\boldsymbol{\beta}^*\|_1 \vee 1) O_p\left(\sqrt{\frac{\log d}{n}}\right) + \|\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2\|_\infty\right] \\
&= o_p(1).
\end{aligned}$$

by assumption.

Similarly one can show that:

$$\begin{aligned}
&|I_X(\mathbf{v}^*, \widehat{\boldsymbol{\beta}}) + I_Y(\mathbf{v}^*, \widehat{\boldsymbol{\beta}}) - I_X(\mathbf{v}^*, \boldsymbol{\beta}^*) - I_Y(\mathbf{v}^*, \boldsymbol{\beta}^*)| \\
&\leq \|\mathbf{v}^*\|_1 \|\boldsymbol{\beta}^*\|_1 (\|\boldsymbol{\beta}^*\|_1 \vee 1) s \log(nd) O_p\left(\sqrt{\frac{\log d}{n}}\right)\left(1 + \|\mathbf{v}^*\|_1 \sqrt{\frac{\log d}{n}}\right) \\
&= o_p(1).
\end{aligned}$$

Define the following notation:

$$\widetilde{M} = \max_{i=1,\ldots,n} \|\boldsymbol{U}_i^{\otimes 2}\|_{\max}.$$

For exactly the same reasons as for M we have $\widetilde{M} = O_p(\log(nd))$. Next we consider the difference:

$$\begin{aligned}
&|I_X(\mathbf{v}^*, \boldsymbol{\beta}^*) + I_Y(\mathbf{v}^*, \boldsymbol{\beta}^*) - \frac{1}{n}\sum_{i=1}^n \left(\mathbf{v}^{*T} \boldsymbol{U}_i^{\otimes 2} \boldsymbol{\beta}^*\right)^2| \\
&\leq \|\mathbf{v}^*\|_1 \|\boldsymbol{\beta}^*\|_1 V \left(\frac{1}{n}\sum_{i=1}^{n_1} |\mathbf{v}^{*T}(\boldsymbol{X}_i - \bar{\boldsymbol{X}})||(\boldsymbol{X}_i - \bar{\boldsymbol{X}})^T \boldsymbol{\beta}^*| + \frac{1}{n}\sum_{i=1}^{n_1} |\mathbf{v}^{*T}(\boldsymbol{Y}_i - \bar{\boldsymbol{Y}})||(\boldsymbol{Y}_i - \bar{\boldsymbol{Y}})^T \boldsymbol{\beta}^*| \right. \\
&\qquad \left. + \frac{1}{n}\sum_{i=1}^n |\mathbf{v}^{*T} \boldsymbol{U}_i||\boldsymbol{U}_i^T \boldsymbol{\beta}^*|\right),
\end{aligned}$$

where:

$$V = \max\left\{\max_{i=1,\ldots,n_1} \|(\boldsymbol{X}_i - \bar{\boldsymbol{X}})^{\otimes 2} - \boldsymbol{U}_i^{\otimes 2}\|_{\max}, \max_{i=n_1+1,\ldots,n} \|(\boldsymbol{Y}_i - \bar{\boldsymbol{Y}})^{\otimes 2} - \boldsymbol{U}_i^{\otimes 2}\|_{\max}\right\}.$$

Note that by the simple inequality $|ab| \leq (a^2 + b^2)/2$, we have, that the expression in the brackets is bounded by:

$$\leq \mathbf{v}^{*T}(\widehat{\boldsymbol{\Sigma}}_n + \widetilde{\boldsymbol{\Sigma}}_n)\mathbf{v}^*/2 + \boldsymbol{\beta}^{*T}(\widehat{\boldsymbol{\Sigma}}_n + \widetilde{\boldsymbol{\Sigma}}_n)\boldsymbol{\beta}^*/2.$$

We have that $\mathbf{v}^{*T}\widehat{\boldsymbol{\Sigma}}_n\mathbf{v}^* \leq \|\mathbf{v}^*\|_1\|\mathbf{v}^{*T}\widehat{\boldsymbol{\Sigma}}_n\|_\infty = \|\mathbf{v}^*\|_1 + \|\mathbf{v}^*\|_1^2 O_p\left(\sqrt{\frac{\log d}{n}}\right)$. Similarly since by (C.4.3) $\|\widehat{\boldsymbol{\Sigma}}_n - \widetilde{\boldsymbol{\Sigma}}_n\|_{\max} = O_p\left(\frac{\log d}{n}\right)$ we have that $\mathbf{v}^{*T}\widetilde{\boldsymbol{\Sigma}}_n\mathbf{v}^* \leq \|\mathbf{v}^*\|_1 + \|\mathbf{v}^*\|_1^2 O_p\left(\sqrt{\frac{\log d}{n}}\right)$. Similarly one can show that $\boldsymbol{\beta}^{*T}\widehat{\boldsymbol{\Sigma}}_n\boldsymbol{\beta}^* \leq \|\boldsymbol{\beta}^*\|_1\|\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2\|_\infty + \|\boldsymbol{\beta}^*\|_1(\|\boldsymbol{\beta}^*\|_1 \vee 1)O_p\left(\sqrt{\frac{\log d}{n}}\right)$, and a similar inequality for $\boldsymbol{\beta}^{*T}\widetilde{\boldsymbol{\Sigma}}_n\boldsymbol{\beta}^*$.

We next inspect V:

$$\begin{aligned}\max_{i=1,\ldots,n_1} \|(\boldsymbol{X}_i - \bar{\boldsymbol{X}})^{\otimes 2} - \boldsymbol{U}_i^{\otimes 2}\|_{\max} \leq \max_{i=1,\ldots,n_1} & 2\|\boldsymbol{X}_i\|_\infty\|\bar{\boldsymbol{X}} - \boldsymbol{\mu}_1\|_\infty \\ & + \|\bar{\boldsymbol{X}} - \boldsymbol{\mu}_1\|_\infty(\|\bar{\boldsymbol{X}} - \boldsymbol{\mu}_1\|_\infty + 2\|\boldsymbol{\mu}_1\|_\infty),\end{aligned}$$

and we can similarly bound the other term in V. Note that in Lemma C.2.8 we showed that $\max_{i=1,\ldots,n_1}\|\boldsymbol{X}_i\|_\infty = O_p(\sqrt{\log(nd)})$, and as we argue in (C.4.2), we have $\|\bar{\boldsymbol{X}} - \boldsymbol{\mu}_1\|_\infty = O_p\left(\sqrt{\frac{\log d}{n}}\right)$, and thus:

$$V = O_p\left(\sqrt{\frac{\log d}{n}}\right)(\sqrt{\log(nd)} + \|\boldsymbol{\mu}_1\|_\infty + \|\boldsymbol{\mu}_2\|_\infty).$$

Hence under our assumptions, we have:

$$|I_X(\mathbf{v}^*, \boldsymbol{\beta}^*) + I_Y(\mathbf{v}^*, \boldsymbol{\beta}^*) - \frac{1}{n}\sum_{i=1}^n \left(\mathbf{v}^{*T}\boldsymbol{U}_i^{\otimes 2}\boldsymbol{\beta}^*\right)^2| = o_p(1).$$

Finally we finish this part upon noting that:

$$\left| \frac{1}{n} \sum_{i=1}^{n} \left(\mathbf{v}^{*T} \boldsymbol{U}_i^{\otimes 2} \boldsymbol{\beta}^* \right)^2 - \mathbb{E} \left(\mathbf{v}^{*T} \boldsymbol{U}_i^{\otimes 2} \boldsymbol{\beta}^* \right)^2 \right| = o_p(1).$$

Under the assumption $\mathrm{Var}((\mathbf{v}^T * \boldsymbol{U}^{\otimes 2} \boldsymbol{\beta}^*)^2) = o(n)$ by Chebyshev's inequality.

Next we turn our attention to the term:

$$\frac{n}{n_1} \frac{1}{n_1} \sum_{i=1}^{n_1} \left(\widehat{\mathbf{v}}^T (\boldsymbol{X}_i - \bar{\boldsymbol{X}}) \right)^2$$

and show it's consistent for $\alpha^{-1} \mathbb{E} (\mathbf{v}^{*T} \boldsymbol{U}_i)^2$. First note that since $\frac{n}{n_1} = \alpha^{-1} + o(\frac{1}{n})$, and we will show the rest of the expression is $O_p(1)$, we will just focus on the average term. We first show the following difference is small:

$$\begin{aligned} \left| \frac{1}{n_1} \sum_{i=1}^{n_1} [\left(\widehat{\mathbf{v}}^T (\boldsymbol{X}_i - \bar{\boldsymbol{X}}) \right)^2 - \left(\mathbf{v}^{*T} (\boldsymbol{X}_i - \bar{\boldsymbol{X}}) \right)^2] \right| &= | \left(\widehat{\mathbf{v}} - \mathbf{v}^* \right)^T \widehat{\boldsymbol{\Sigma}}_{\boldsymbol{X}} \left(\widehat{\mathbf{v}} + \mathbf{v}^* \right)^T | \\ &\leq \|\widehat{\mathbf{v}} - \mathbf{v}^*\|_1 (\|\widehat{\mathbf{v}} - \mathbf{v}^*\|_1 + 2\|\mathbf{v}^*\|_1) \|\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{X}}\|_\infty. \end{aligned}$$

Using the same technique as in the proof of Lemma C.4.1, one can show that $\|\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{X}}\|_\infty \leq \|\boldsymbol{\Sigma}\|_\infty + O_p\left(\sqrt{\frac{\log d}{n}}\right)$. Since we also know by Lemma C.4.6 that $\|\widehat{\mathbf{v}} - \mathbf{v}^*\|_1 = \|\mathbf{v}^*\|_1 s_{\mathbf{v}} O_p\left(\sqrt{\frac{\log d}{n}}\right)$, we get:

$$\frac{1}{n_1} \left| \sum_{i=1}^{n_1} [\left(\widehat{\mathbf{v}}^T (\boldsymbol{X}_i - \bar{\boldsymbol{X}}) \right)^2 - \left(\mathbf{v}^{*T} (\boldsymbol{X}_i - \bar{\boldsymbol{X}}) \right)^2] \right| \leq \|\mathbf{v}^*\|_1^2 s_{\mathbf{v}} O_p\left(\sqrt{\frac{\log d}{n}}\right) = o_p(1),$$

by assumption. Next we control:

$$\begin{aligned}\left|\frac{1}{n_1}\sum_{i=1}^{n_1}[\left(\mathbf{v}^{*T}(\boldsymbol{X}_i-\bar{\boldsymbol{X}})\right)^2-\left(\mathbf{v}^{*T}\boldsymbol{U}_i\right)^2]\right| &= \left|\mathbf{v}^{*T}(\boldsymbol{\mu}_1-\bar{\boldsymbol{X}})\frac{1}{n_1}\sum_{i=1}^{n_1}(2\boldsymbol{X}_i-\bar{\boldsymbol{X}}-\boldsymbol{\mu}_1)\mathbf{v}^*\right| \\ &\leq \|\mathbf{v}^*\|_1^2\|\boldsymbol{\mu}_1-\bar{\boldsymbol{X}}\|_\infty^2 \\ &= \|\mathbf{v}^*\|_1^2 O_p\left(\frac{\log d}{n}\right) \\ &= o_p(1).\end{aligned}$$

Thus after using Chebyshev's inequality upon observing that $\mathrm{Var}((\mathbf{v}^{*T}\boldsymbol{U})^2) = o(n)$, we have shown the desired consistency. Similarly we can also show that $\frac{n}{n_2}\frac{1}{n_2}\sum_{i=n_1+1}^{n}\left(\widehat{\mathbf{v}}^T(\boldsymbol{Y}_i-\bar{\boldsymbol{Y}})\right)^2$ is consistent for $(\alpha-1)^{-1}\mathbb{E}(\mathbf{v}^{*T}\boldsymbol{U}_i)^2$. This concludes the proof.

□

Lemma C.4.1. *The following inequality holds:*

$$\|\widehat{\boldsymbol{\Sigma}}_n-\boldsymbol{\Sigma}\|_{\max}\leq \widetilde{t}_{\boldsymbol{U}}(d,n)+t_{\boldsymbol{U}}^2(d,n).$$

with probability at least $1-2d^{2(1-\widetilde{c}_U A_{\boldsymbol{U}}^2)}-2ed^{1-c_U A_{\boldsymbol{U}}^2}$*, where:*

$$t_{\boldsymbol{U}}(d,n)=A_{\boldsymbol{U}}K_{\boldsymbol{U}}\sqrt{\frac{\log d}{\min(n_1,n_2)}};\quad \widetilde{t}_{\boldsymbol{U}}(d,n)=2A_{\boldsymbol{U}}K_{\boldsymbol{U}}^2\sqrt{\frac{\log d}{n}}.$$

and $A_{\boldsymbol{U}}>0$ *is an arbitrary positive constant,* $\widetilde{c}_{\boldsymbol{U}}$ *and* $c_{\boldsymbol{U}}$ *are absolute contents independent of the distribution of* $\boldsymbol{U}$*, and* $K_{\boldsymbol{U}}$ *is as defined in the main section of the text.*

Proof of Lemma C.4.1. We start by showing a concentration bound on $\|\bar{\boldsymbol{X}}-\boldsymbol{\mu}_1\|_\infty$ and $\|\bar{\boldsymbol{Y}}-$

$\boldsymbol{\mu}_2\|_\infty$. By proposition 5.10 in Vershynin [84] and the union bound, we have:

$$\mathbb{P}(\|\bar{\boldsymbol{X}} - \boldsymbol{\mu}_1\|_\infty > t) \leq ed \exp\left(-\frac{c_{\boldsymbol{U}} n_1 t^2}{K_{\boldsymbol{U}}^2}\right). \tag{C.4.2}$$

A similar inequality holds for $\|\bar{\boldsymbol{Y}} - \boldsymbol{\mu}_2\|_\infty$. Select $t_{\boldsymbol{U}}(d,n) = A_{\boldsymbol{U}} K_{\boldsymbol{U}} \sqrt{\frac{\log d}{\min(n_1,n_2)}}$, where $A_{\boldsymbol{U}} > 0$ is some large constant. The triangle inequality yields:

$$\|\widehat{\boldsymbol{\Sigma}}_n - \boldsymbol{\Sigma}\|_{\max} \leq \|\widehat{\boldsymbol{\Sigma}}_n - \widetilde{\boldsymbol{\Sigma}}_n\|_{\max} + \|\widetilde{\boldsymbol{\Sigma}}_n - \boldsymbol{\Sigma}\|_{\max},$$

where $\widetilde{\boldsymbol{\Sigma}}_n$ is defined as in (C.4.1). Next, we have that:

$$\begin{aligned}
\|\widehat{\boldsymbol{\Sigma}}_n - \widetilde{\boldsymbol{\Sigma}}_n\|_{\max} &\leq \frac{n_1}{n}\|(\bar{\boldsymbol{X}} - \boldsymbol{\mu}_1)^{\otimes 2}\|_{\max} + \frac{n_2}{n}\|(\bar{\boldsymbol{Y}} - \boldsymbol{\mu}_2)^{\otimes 2}\|_{\max} \\
&\leq \frac{n_1}{n}(\|\bar{\boldsymbol{X}} - \boldsymbol{\mu}_1\|_\infty)^2 + \frac{n_2}{n}(\|\bar{\boldsymbol{Y}} - \boldsymbol{\mu}_2\|_\infty)^2 \\
&\leq t_{\boldsymbol{U}}^2(d,n). \qquad \text{(C.4.3)}
\end{aligned}$$

where the last inequality holds with high probability. Note that by Lemma C.2.2 we have:

$$\|\widetilde{\boldsymbol{\Sigma}}_n - \boldsymbol{\Sigma}\|_{\max} \leq 2A_{\boldsymbol{U}} K_{\boldsymbol{U}}^2 \sqrt{\frac{\log d}{n}} =: \widetilde{t}_{\boldsymbol{U}}(d,n),$$

with probability at least $1 - 2d^{2(1-\widetilde{c}_U A_{\boldsymbol{U}}^2)}$. Adding the last two inequalities completes the proof.

□

Lemma C.4.2. *Assume the same conditions as in Lemma C.4.1, and assume further that the minimum eigenvalue* $\lambda_{\min}(\boldsymbol{\Sigma}) > 0$ *and* $s(\widetilde{t}_{\boldsymbol{U}}(d,n) + t_{\boldsymbol{U}}^2(d,n)) \leq (1-\kappa)\frac{\lambda_{\min}(\boldsymbol{\Sigma})}{(1+\xi)^2}$*, where* $0 < \kappa < 1$*. We then have that* $\widehat{\boldsymbol{\Sigma}}_n$ *satisfies the RE property with* $\mathrm{RE}_{\widehat{\boldsymbol{\Sigma}}_n}(s,\xi) \geq \kappa\lambda_{\min}(\boldsymbol{\Sigma})$ *with probability at least* $1 - 2d^{2(1-\widetilde{c}_U A_{\boldsymbol{U}}^2)} - 2ed^{1-c_U A_{\boldsymbol{U}}^2}$.

Remark C.4.3. *In fact this event happens on the same event as in Lemma C.4.1.*

Proof of Lemma C.4.2. The proof follows the proof of Lemma C.2.3, but uses Lemma C.4.1 instead of Lemma C.2.2, hence we omit it. □

Lemma C.4.4. *Assume that* — $\lambda_{\min}(\mathbf{\Sigma}) > 0$, $s(\widetilde{t}_{\boldsymbol{U}}(d,n) + t^2_{\boldsymbol{U}}(d,n)) \leq (1-\kappa)\frac{\lambda_{\min}(\mathbf{\Sigma})}{(1+\xi)^2}$, *where* $0 < \kappa < 1$ *and:*

$$\lambda \geq \left(\widetilde{t}_{\boldsymbol{U}}(d,n) + t^2_{\boldsymbol{U}}(d,n)\right) \|\boldsymbol{\beta}^*\|_1 + 2t_{\boldsymbol{U}}(d,n).$$

Then we have that $\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_1 \leq \frac{8\lambda s}{\mathrm{RE}(s,1)}$ *with probability at least* $1 - 2d^{2(1-\widetilde{c}_{\boldsymbol{U}}A^2_{\boldsymbol{U}})} - 2ed^{1-c_{\boldsymbol{U}}A^2_{\boldsymbol{U}}}$.

Remark C.4.5. *In fact this event happens on the same event as in Lemma C.4.1.*

Proof of Lemma C.4.4. We start by showing the true parameter $\mathbf{\Omega}\boldsymbol{\delta} = \boldsymbol{\beta}^*$ satisfies the sparse LDA constraint — $\|\widehat{\mathbf{\Sigma}}_n\boldsymbol{\beta}^* - (\bar{\boldsymbol{X}} - \bar{\boldsymbol{Y}})\|_\infty \leq \lambda$ with probability at least $1 - 2d^{2(1-\widetilde{c}_{\boldsymbol{U}}A^2_{\boldsymbol{U}})} - 2ed^{1-c_{\boldsymbol{U}}A^2_{\boldsymbol{U}}}$. We have that:

$$\|\widehat{\mathbf{\Sigma}}_n\boldsymbol{\beta}^* - (\bar{\boldsymbol{X}} - \bar{\boldsymbol{Y}})\|_\infty \leq \underbrace{\|\mathbf{\Sigma}\boldsymbol{\beta}^* - (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)\|_\infty}_{0} + \|\widehat{\mathbf{\Sigma}}_n - \mathbf{\Sigma}\|_{\max}\|\boldsymbol{\beta}^*\|_1 + \|\bar{\boldsymbol{X}} - \boldsymbol{\mu}_1\|_\infty + \|\bar{\boldsymbol{Y}} - \boldsymbol{\mu}_2\|_\infty.$$

Collecting the bounds we derived in Lemma C.4.1 we get:

$$\|\widehat{\mathbf{\Sigma}}_n\boldsymbol{\beta}^* - (\bar{\boldsymbol{X}} - \bar{\boldsymbol{Y}})\|_\infty \leq \left(\widetilde{t}_{\boldsymbol{U}}(d,n) + t^2_{\boldsymbol{U}}(d,n)\right) \|\boldsymbol{\beta}\|_1^* + 2t_{\boldsymbol{U}}(d,n).$$

The last inequality implies that if we select $\lambda \geq \left(\widetilde{t}_{\boldsymbol{U}}(d,n) + t^2_{\boldsymbol{U}}(d,n)\right) \|\boldsymbol{\beta}\|_1^* + 2t_{\boldsymbol{U}}(d,n)$, it will follow that $\boldsymbol{\beta}^*$ satisfies the constraint with probability at least $1 - 2d^{2(1-c_{\boldsymbol{U}}A^2_{\boldsymbol{U}})} - 2ed^{1-c_{\boldsymbol{U}}A^2_{\boldsymbol{U}}}$.

The rest of the proof is identical to the proof of Lemma C.2.5 but instead of using Lemma C.2.3 we use Lemma C.4.2. Thus we omit the proof. □

Lemma C.4.6. *Assume that* $-\lambda_{\min}(\mathbf{\Sigma}) > 0$, $s_{\mathbf{v}}(\widetilde{t}_{\boldsymbol{U}}(d,n) + t^2_{\boldsymbol{U}}(d,n)) \leq (1-\kappa)\frac{\lambda_{\min}(\mathbf{\Sigma})}{(1+\xi)^2}$, *where* $0 < \kappa < 1$ *and* $\lambda' \geq \|\mathbf{v}^*\|_1(\widetilde{t}_{\boldsymbol{U}}(d,n) + t^2_{\boldsymbol{U}}(d,n))$. *Then we have that* $\|\widehat{\mathbf{v}} - \mathbf{v}^*\|_1 \leq \frac{8\lambda' s_{\mathbf{v}}}{\mathrm{RE}_\kappa(s_{\mathbf{v}},1)}$ *with probability at least* $1 - 2d^{2(1-\widetilde{c}_U A^2_{\boldsymbol{U}})} - 2ed^{1-c_U A^2_{\boldsymbol{U}}}$.

Remark C.4.7. *In fact this event happens on the same event as in Lemma C.4.1.*

Proof of Lemma C.4.6. The proof is identical to the one of Lemma C.2.5 but instead of using Lemma C.2.3 we use Lemma C.4.2, and we use Lemma C.4.1 instead of using Lemma C.2.2. We omit the proof. □

Lemma C.4.8. *We have the following inequality:*

$$\mathbb{E}|\mathbf{v}^{*T}\boldsymbol{U}^{\otimes 2}\boldsymbol{\beta}^* - \mathbf{v}^{*T}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) + c\mathbf{v}^{*T}\boldsymbol{U}|^3 \leq 4\|\mathbf{v}^*\|_2^3(\|\boldsymbol{\beta}^*\|_2^3(s_{\mathbf{v}}s)^{3/2}(24K^2_{\boldsymbol{U}})^3 + |c|^3 s_{\mathbf{v}}^{3/2}(\sqrt{6}K_{\boldsymbol{U}})^3).$$

Proof of Lemma C.4.8. First note that:

$$\boldsymbol{U}^{\otimes 2}\boldsymbol{\beta}^* - (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) = (\boldsymbol{U}^{\otimes 2} - \mathbf{\Sigma})\boldsymbol{\beta}^*.$$

Denote with $\boldsymbol{\xi} = (\boldsymbol{U}^{\otimes 2} - \mathbf{\Sigma})\boldsymbol{\beta}^*$ for brevity. We have the following by Cauchy-Schwartz and Jensen's inequalities:

$$\mathbb{E}|\mathbf{v}^{*T}\boldsymbol{\xi} + c\mathbf{v}^{*T}\boldsymbol{U}|^3 \leq \|\mathbf{v}^*\|_2^3\sqrt{\mathbb{E}\|\boldsymbol{\xi}_{S_{\mathbf{v}}} + c\boldsymbol{U}_{S_{\mathbf{v}}}\|_2^6}.$$

Using the triangle inequality and the fact that $(a+b)^2 \leq 2(a^2+b^2)$ we have:

$$\|\mathbf{v}^*\|_2^3\sqrt{\mathbb{E}\|\boldsymbol{\xi}_{S_{\mathbf{v}}} + c\boldsymbol{U}_{S_{\mathbf{v}}}\|_2^6} \leq \|\mathbf{v}^*\|_2^3\sqrt{8\mathbb{E}(\|\boldsymbol{\xi}_{S_{\mathbf{v}}}\|_2^2 + c^2\|\boldsymbol{U}_{S_{\mathbf{v}}}\|_2^2)^3}.$$

Next, using the fact that $\left(\frac{a^2+b^2}{2}\right)^{1/2} \leq \left(\frac{a^3+b^3}{2}\right)^{1/3}$, we get

$$\|\mathbf{v}^*\|_2^3\sqrt{8\mathbb{E}(\|\boldsymbol{\xi}_{S_\mathbf{v}}\|_2^2 + c^2\|\boldsymbol{U}_{S_\mathbf{v}}\|_2^2)^3} \leq \|\mathbf{v}^*\|_2^3\sqrt{16\mathbb{E}(\|\boldsymbol{\xi}_{S_\mathbf{v}}\|_2^3 + |c|^3\|\boldsymbol{U}_{S_\mathbf{v}}\|_2^3)^2}.$$

Finally by a triangle inequality we have:

$$\|\mathbf{v}^*\|_2^3\sqrt{16\mathbb{E}(\|\boldsymbol{\xi}_{S_\mathbf{v}}\|_2^3 + |c|^3\|\boldsymbol{U}_{S_\mathbf{v}}\|_2^3)^2} \leq 4\|\mathbf{v}^*\|_2^3\left(\sqrt{\mathbb{E}\|\boldsymbol{\xi}_{S_\mathbf{v}}\|_2^6} + |c|^3\sqrt{\mathbb{E}\|\boldsymbol{U}_{S_\mathbf{v}}\|_2^6}\right).$$

Now recall that:

$$\sqrt{\mathbb{E}\|\boldsymbol{\xi}_{S_\mathbf{v}}\|_2^6} = \sqrt{\mathbb{E}\|[(\boldsymbol{U}^{\otimes 2} - \boldsymbol{\Sigma})\boldsymbol{\beta}^*]_{S_\mathbf{v}}\|_2^6} \leq \|\boldsymbol{\beta}^*\|_2^3\sqrt{\mathbb{E}\|\operatorname{Vec}[(\boldsymbol{U}^{\otimes 2} - \boldsymbol{\Sigma})_{S_\mathbf{v},S}]\|_2^6}.$$

Using (C.3.2) and (C.3.3) from Lemma C.3.2, we get:

$$\sqrt{\mathbb{E}\|\boldsymbol{\xi}_{S_\mathbf{v}}\|_2^6} \leq \|\boldsymbol{\beta}^*\|_2^3(s_\mathbf{v}s)^{3/2}(24K_{\boldsymbol{U}}^2)^3.$$

Finally we deal with, using the same argument as in (C.2.12):

$$\sqrt{\mathbb{E}\|\boldsymbol{U}_{S_\mathbf{v}}\|_2^6} \leq s_\mathbf{v}^{3/2}(\sqrt{6}K_{\boldsymbol{U}})^3.$$

Putting everything together, we get:

$$\mathbb{E}|\mathbf{v}^{*T}\boldsymbol{\xi} + c\mathbf{v}^{*T}\boldsymbol{U}|^3 \leq 4\|\mathbf{v}^*\|_2^3(\|\boldsymbol{\beta}^*\|_2^3(s_\mathbf{v}s)^{3/2}(24K_{\boldsymbol{U}}^2)^3 + |c|^3 s_\mathbf{v}^{3/2}(\sqrt{6}K_{\boldsymbol{U}})^3),$$

as claimed. □

C.5 Proofs for SVA

Proof of Theorem 4.7.1. Let $\widehat{\boldsymbol{\beta}}_0 = (0, \widehat{\boldsymbol{\gamma}}^T)^T$. Note that the proof of Theorem 4.3.3 extends in this case, as it does not rely on the iid representation. Using Theorem C.5.1 we have $\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_1 \leq 4s\|\Sigma_0^{-1}\|_1\lambda$. Furthermore $\|S_0\boldsymbol{\beta}^* - S_{1,*m}\|_\infty \leq \lambda$ with probability at least $1 - 14d^{-1}$, as can be seen in the proof of Theorem C.5.1. Next we inspect $\|\widehat{\mathbf{v}} - \mathbf{v}^*\|_1 = O_p(s_{\mathbf{v}}\|\Sigma_0^{-1}\|_1\lambda')$ according to Lemma C.5.2. Furthermore, Lemma C.5.2 also gives us that $\|\widehat{[\mathbf{v}^T S_0]_{-1}}\|_\infty = O_p(\lambda')$. Combining these results with the assumptions from the statement completes the proof. □

Proof of Theorem 4.7.2. First, construct the following sequence $\xi_1 = 0, \xi_{t+1} = \frac{\mathbf{v}^{*T}\boldsymbol{X}_t^{\otimes 2}\boldsymbol{\beta}^* - \mathbf{v}^{*T}\boldsymbol{X}_t\boldsymbol{X}_{t+1}^T\mathbf{e}_m^T}{\sqrt{(T-1)\boldsymbol{\Psi}_{mm}\mathbf{v}^{*T}\Sigma_0\mathbf{v}^*}}$ for $t = 1, \ldots, T-1$. We start by showing that the difference between the sequence $\sum_{t=1}^T \xi_t$ and $\frac{\sqrt{T-1}\mathbf{v}^{*T}(S_0\boldsymbol{\beta}^* - S_{1,*m})}{\sqrt{\boldsymbol{\Psi}_{mm}\mathbf{v}^{*T}\Sigma_0\mathbf{v}^*}}$ is asymptotically negligible. We have:

$$\left|\sum_{t=1}^T \xi_t - \frac{\sqrt{T-1}\mathbf{v}^{*T}(S_0\boldsymbol{\beta}^* - S_{1,*m})}{\sqrt{\boldsymbol{\Psi}_{mm}\mathbf{v}^{*T}\Sigma_0\mathbf{v}^*}}\right| \leq \underbrace{\frac{(\sqrt{T-1}T)^{-1}}{\sqrt{\boldsymbol{\Psi}_{mm}\mathbf{v}^{*T}\Sigma_0\mathbf{v}^*}}\left|\sum_{t=1}^T \mathbf{v}^{*T}\boldsymbol{X}_t^{\otimes 2}\boldsymbol{\beta}^*\right|}_{I_1} + \underbrace{\frac{|\mathbf{v}^{*T}\boldsymbol{X}_T^{\otimes 2}\boldsymbol{\beta}^*|}{\sqrt{T-1}\sqrt{\boldsymbol{\Psi}_{mm}\mathbf{v}^{*T}\Sigma_0\mathbf{v}^*}}}_{I_2}.$$

By Lemma C.5.2 $\|\mathbf{v}^{*T}S_0 - \mathbf{e}\|_\infty \leq \lambda'$ with probability not smaller than $1 - 14d^{-1}$. Thus under the null hypothesis we have:

$$I_1 \leq \frac{(\sqrt{T-1})^{-1}\lambda'\|\boldsymbol{\beta}^*\|_1}{\sqrt{\boldsymbol{\Psi}_{mm}\mathbf{v}^{*T}\Sigma_0\mathbf{v}^*}} \leq \frac{(\sqrt{T-1})^{-1}\lambda' M_d}{\sqrt{\boldsymbol{\Psi}_{mm}\mathbf{v}^{*T}\Sigma_0\mathbf{v}^*}} = o(1),$$

with probability at least $1 - 14d^{-1}$. Next for I_2 we have:

$$I_2 = \frac{|Z_1||Z_2|}{\sqrt{T-1}},$$

where $Z_1 \sim N(0,1)$ and $Z_2 \sim N(0, \frac{\boldsymbol{\beta}^{*T}\Sigma_0\boldsymbol{\beta}^*}{\boldsymbol{\Psi}_{mm}})$, and hence since $\frac{\boldsymbol{\beta}^{*T}\Sigma_0\boldsymbol{\beta}^*}{\boldsymbol{\Psi}_{mm}} = o(T)$, Chebyshev's

inequality gives $I_2 = o_p(1)$. Of course by Slutsky's theorem the last implies that:

$$\left| \sum_{t=1}^{T} \xi_t - \frac{\sqrt{T}\mathbf{v}^{*T}(S_0\boldsymbol{\beta}^* - S_{1,*m})}{\sqrt{\boldsymbol{\Psi}_{mm}\mathbf{v}^{*T}\Sigma_0\mathbf{v}^*}} \right| = o_p(1),$$

provided that $\sum_{t=1}^{T} \xi_t = O_p(1)$, which we show next.

Obverve that the sequence $(\xi_t)_{t=1}^{T}$ forms a martingale difference sequence with respect to the filtration $\mathcal{F}_t = \sigma(\boldsymbol{X}_1, \ldots, \boldsymbol{X}_t)$ for $t = 1, \ldots, T$, as we clearly have $\mathbb{E}[\xi_t|\mathcal{F}_{t-1}] = 0$. Furthermore a simple calculation yields that for $t \geq 2$ we have $\mathbb{E}[\xi_t^2|\mathcal{F}_{t-1}] = \frac{(\mathbf{v}^{*T}\boldsymbol{X}_{t-1})^2}{(T-1)\mathbf{v}^{*T}\Sigma_0\mathbf{v}^*}$. Thus:

$$\left| \sum_{t=1}^{T} \mathbb{E}[\xi_t^2|\mathcal{F}_{t-1}] - 1 \right| = \left| \frac{\mathbf{v}^{*T}}{(T-1)\mathbf{v}^{*T}\Sigma_0\mathbf{v}^*} \sum_{t=1}^{T-1} \left[\boldsymbol{X}_t^{\otimes 2} - \Sigma_0\right] \mathbf{v}^* \right| \leq \frac{\|\mathbf{v}^*\|_1^2}{\mathbf{v}^{*T}\Sigma_0\mathbf{v}^*} \underbrace{\left\| \frac{1}{T-1} \sum_{t=1}^{T-1} [\boldsymbol{X}_t^{\otimes 2} - \Sigma_0] \right\|_{\max}}_{I}.$$

Using Theorem C.5.1, it is evident that $I \leq K_d(\Sigma_0, A)/2 \left(\sqrt{\frac{6 \log d}{T-1}} + 2\sqrt{\frac{1}{T-1}} \right)$ with probability at least $1 - 14d^{-1}$, and hence the above quantity converges to 0 in probability.

Having noted these facts, we want to show that $\sum_{t=1}^{T} \xi_t$ converges weakly to a $N(0,1)$ with the help of a version of the martingale central limit theorem (MCLT)[28]. Next we show the Lindeberg condition for the MCLT. For $t \geq 2$ and a fixed $\delta > 0$ we have:

$$\mathbb{E}[\xi_t^2 1(|\xi_t| \geq \delta)|\mathcal{F}_{t-1}] = \frac{(\mathbf{v}^{*T}\boldsymbol{X}_{t-1})^2 \mathbb{E}[Z^2 1(|Z| > \delta C)]}{(T-1)\mathbf{v}^{*T}\Sigma_0\mathbf{v}^*},$$

where $Z \sim N(0,1)$ and $C = \left\{ \frac{(\mathbf{v}^{*T}\boldsymbol{X}_{t-1})^2}{(T-1)\mathbf{v}^{*T}\Sigma_0\mathbf{v}^*} \right\}^{-\frac{1}{2}}$. Using the properties of the truncated standard normal distribution we have that $\mathbb{E}[Z^2|Z > c] = 1 + \frac{\phi(c)}{\overline{\Phi}(c)}c$, and hence

$$\mathbb{E}[Z^2 1(|Z| > c)] = 2\overline{\Phi}(c)\left(1 + \frac{\phi(c)}{\overline{\Phi}(c)}c\right) = 2\overline{\Phi}(c) + 2\phi(c)c \leq 2\phi(c)(c^{-1} + c),$$

where the last inequality follows from a standard tail bound for the normal distribution.

Now notice that by the union bound and a standard bound on the normal cdf we have

$$\mathbb{P}(\max_{t=1,\ldots,T} |\mathbf{v}^{*T}\boldsymbol{X}_t| > u) \leq 2T\exp(-u^2/(2\mathbf{v}^{*T}\Sigma_0\mathbf{v}^*)).$$

Selecting $u = 2\sqrt{\log(T)\mathbf{v}^{*T}\Sigma_0\mathbf{v}^*}$ gives that with probability at least $1-\frac{2}{T}$ we have $\max_t |\mathbf{v}^{*T}\boldsymbol{X}_t| \leq 2\sqrt{\log(T)\mathbf{v}^{*T}\Sigma_0\mathbf{v}^*}$. Hence on this event we have:

$$\mathbb{E}[\xi_t^2 1(|\xi_t| \geq \delta)|\mathcal{F}_{t-1}] \leq \frac{8\log T\phi(\delta\widetilde{C})((\delta\widetilde{C})^{-1} + \delta\widetilde{C})}{(T-1)},$$

where $\widetilde{C} = \sqrt{\frac{T-1}{4\log T}}$, and we used the fact that the function $\phi(x)(x^{-1}+x)$ is decreasing. Summing up over t yields:

$$\sum_{i=1}^{T} \mathbb{E}[\xi_t^2 1(|\xi_t| \geq \delta)|\mathcal{F}_{t-1}] \leq 8\log T\phi(\delta\widetilde{C})((\delta\widetilde{C})^{-1} + \delta\widetilde{C}) \to 0.$$

This shows that the Lindeberg condition holds with probability 1. Hence by the MCLT we can claim:

$$\sum_{t=1}^{T} \xi_t \rightsquigarrow N(0,1).$$

□

Proof of Proposition 4.7.3. We begin with showing the consistency of $\widehat{\boldsymbol{\Psi}}_{mm} = S_{0,mm} - \widehat{\boldsymbol{\beta}}^T S_0 \widehat{\boldsymbol{\beta}}$ is consistent for $\boldsymbol{\Psi}_{mm}$. First note that $\boldsymbol{\Psi} = \Sigma_0 - A^T\Sigma_0 A$, and thus $\boldsymbol{\Psi}_{mm} = \Sigma_{0,mm} - \boldsymbol{\beta}^{*T}\Sigma_0\boldsymbol{\beta}^*$. Then we have:

$$|\widehat{\boldsymbol{\Psi}}_{mm} - \boldsymbol{\Psi}_{mm}| \leq |S_{0,mm} - \Sigma_{0,mm}| + |(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)^T S_0\widehat{\boldsymbol{\beta}}| + |\boldsymbol{\beta}^{*T}(S_0\widehat{\boldsymbol{\beta}} - \Sigma_0\boldsymbol{\beta}^*)|.$$

Fitstly, by Theorem C.5.1, we have with probability at least $1 - 14d^{-1}$:

$$|S_{0,mm} - \Sigma_{0,mm}| \leq \|S_0 - \Sigma_0\|_{\max} \leq K_d(\Sigma_0, A)/2 \left(\sqrt{\frac{6 \log d}{T}} + 2\sqrt{\frac{1}{T}} \right) = \lambda' \|\Sigma_0^{-1}\|_1^{-1} = o(1).$$

Secondly:

$$\begin{aligned} |(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)^T S_0 \widehat{\boldsymbol{\beta}}| &\leq \|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_1 (\|S_0 \boldsymbol{\beta}^*\|_\infty + \|S_0 \widehat{\boldsymbol{\beta}} - S_0 \boldsymbol{\beta}^*\|_\infty) \\ &\leq \|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_1 (\|S_0\|_{\max} \|\boldsymbol{\beta}^*\|_1 + \|S_0 \widehat{\boldsymbol{\beta}} - S_{1,*m}\|_\infty + \|S_0 \boldsymbol{\beta}^* - S_{1,*m}\|_\infty). \end{aligned}$$

On the event of Theorem C.5.1 we further have:

$$\|\boldsymbol{\beta} - \boldsymbol{\beta}^*\|_1 \|S_0\|_{\max} \|\boldsymbol{\beta}^*\|_1 \leq 4s \|\Sigma_0^{-1}\|_1 \lambda [\|\Sigma_0\|_{\max} + \|S_0 - \Sigma_0\|_{\max}] M_d = o(1).$$

Furthermore within the proof of Theorem C.5.1, it can be seen that on the event of interest we have $\|S_0 \boldsymbol{\beta}^* - S_{1,*m}\|_\infty \leq \lambda$, and hence:

$$\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_1 (\|S_0 \widehat{\boldsymbol{\beta}} - S_{1,*m}\|_\infty + \|S_0 \boldsymbol{\beta}^* - S_{1,*m}\|_\infty) \leq 2\lambda \|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_1 = o_p(1).$$

Lastly,

$$\begin{aligned} |\boldsymbol{\beta}^{*T} (S_0 \widehat{\boldsymbol{\beta}} - \Sigma_0 \boldsymbol{\beta}^*)| &\leq |\boldsymbol{\beta}^{*T} S_0 (\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)| + |\boldsymbol{\beta}^{*T} (S_0 - \Sigma_0) \boldsymbol{\beta}^*| \\ &\leq \|\boldsymbol{\beta}^*\|_1 2\lambda + \|\boldsymbol{\beta}^*\|_1 [\|S_0 \boldsymbol{\beta}^* - S_{1,m}\|_{\max} + \|S_{1,*m} - \Sigma_{1,*m}\|_{\max}] \\ &\leq M_d 3\lambda + M_d K_d(\Sigma_0, A) \left(\sqrt{\frac{3 \log d}{T}} + \sqrt{\frac{2}{T}} \right) = o(1), \end{aligned}$$

where the last two inequalities hold on the event of Theorem C.5.1, and we used the fact that $\|\boldsymbol{\beta}^*\|_1 \leq M_d$ since $A \in \mathcal{M}(s, M_d)$.

Next, we show that $\widehat{\mathbf{v}}^T S_0 \widehat{\mathbf{v}} \to_p \mathbf{v}^{*T} \Sigma_0 \mathbf{v}^*$. Similarly to before we have:

$$|\widehat{\mathbf{v}}^T S_0 \widehat{\mathbf{v}} - \mathbf{v}^* \Sigma_0 \mathbf{v}^*| \leq |(\widehat{\mathbf{v}} - \mathbf{v}^*)^T S_0 \widehat{\mathbf{v}}| + |\mathbf{v}^{*T}(S_0 \widehat{\mathbf{v}} - \Sigma_0 \mathbf{v}^*)|$$

For the firs termt we have:

$$\begin{aligned} |(\widehat{\mathbf{v}} - \mathbf{v}^*) S_0 \widehat{\mathbf{v}}| &\leq \|\widehat{\mathbf{v}} - \mathbf{v}^*\|_1 \|\widehat{\mathbf{v}}^T S_0 - \mathbf{e}\|_\infty + |\mathbf{e}(\widehat{\mathbf{v}} - \mathbf{v}^*)| \\ &\leq 4 s_{\mathbf{v}} \|\Sigma_0^{-1}\|_1 (\lambda')^2 + \|\widehat{\mathbf{v}} - \mathbf{v}^*\|_\infty \\ &\leq 4 s_{\mathbf{v}} \|\Sigma_0^{-1}\|_1 (\lambda')^2 + \|\Sigma_0^{-1}\|_1 2\lambda' = o(1), \end{aligned}$$

with the last two inequalities following from Lemma C.5.2 and holding on the event from Theorem C.5.1. Recall that $\mathbf{e}$ is a unit row vector.

Finally, for the second term we have:

$$|\mathbf{v}^{*T}(S_0 \widehat{\mathbf{v}} - \Sigma_0 \mathbf{v}^*)| \leq \|\mathbf{v}^*\|_1 \lambda' \leq \|\Sigma_0^{-1}\|_1 \lambda' = o(1),$$

and this concludes the proof. □

Theorem C.5.1 (Theorem 4.1.[29]). *Suppose that* $(\boldsymbol{X}_t)_{t=1}^T$ *from a lag* 1 *vector autoregressive process* $(\boldsymbol{X}_t)_{t=-\infty}^{\infty}$. *Assume that* $A \in \mathcal{M}(s, M_d)$. *Let* $\widehat{A}$ *is the optimum of (4.7.1) with the tunning parameter:*

$$\lambda = \widetilde{K}_d(\Sigma_0, A) \sqrt{\frac{\log d}{T}}.$$

For $T \geq 6 \log d + 1$ *and* $d \geq 8$, *we have, with probability at least* $1 - 14 d^{-1}$*:*

$$\|\widehat{A} - A\|_1 \leq 4 s \|\Sigma_0^{-1}\|_1 \lambda.$$

In fact on the same event (see Lemmas A.1. and A.2.[29]), we have:

$$\|S_0-\Sigma_0\|_{\max} \leq K_d(\Sigma_0, A)/2\left(\sqrt{\frac{6\log d}{T}}+2\sqrt{\frac{1}{T}}\right), \quad \|S_1-\Sigma_1\|_{\max} \leq K_d(\Sigma_0, A)\left(\sqrt{\frac{3\log d}{T}}+\sqrt{\frac{2}{T}}\right)$$

Lemma C.5.2. *Assume the assumptions in Theorem C.5.1. Let* $\lambda' = \|\Sigma_0^{-1}\|_1 K_d(\Sigma_0, A)/2\left(\sqrt{\frac{6\log d}{T}}+2\sqrt{\frac{1}{T}}\right)$. *Then on the same event as in Theorem C.5.1, we have* $\|\widehat{\mathbf{v}}-\mathbf{v}^*\|_1 \leq 4s_{\mathbf{v}}\|\Sigma_0^{-1}\|_1\lambda'$.

Proof of Lemma C.5.2. We first start by showing that $\mathbf{v}^*$ satisfies the constraint in the $\widehat{\mathbf{v}}$ optimization problem with high probability. According to Theorem C.5.1, we have with not smaller than $1-14d^{-1}$:

$$\begin{aligned}\|\mathbf{v}^{*T}S_0-\mathbf{e}\|_\infty &= \|\mathbf{v}^{*T}(S_0-\Sigma_0)\|_\infty \leq \|\mathbf{v}^*\|_1\|S_0-\Sigma_0\|_{\max}\\ &\leq \|\mathbf{v}^*\|_1 K_d(\Sigma_0, A)/2\left(\sqrt{\frac{6\log d}{T}}+2\sqrt{\frac{1}{T}}\right) \leq \lambda'.\end{aligned}$$

This implies that $\|\widehat{\mathbf{v}}\|_1 \leq \|\mathbf{v}^*\|_1 \leq \|\Sigma_0^{-1}\|_1$, and hence similarly to (C.2.6) in Lemma C.2.5 we can conclude:

$$\|\widehat{\mathbf{v}}_{S_{\mathbf{v}}^c}-\mathbf{v}^*_{S_{\mathbf{v}}^c}\|_1 \leq \|\widehat{\mathbf{v}}_{S_{\mathbf{v}}}-\mathbf{v}^*_{S_{\mathbf{v}}}\|_1 \tag{C.5.1}$$

Next we control $\|\widehat{\mathbf{v}}-\mathbf{v}^*\|_\infty$. We have:

$$\begin{aligned}\|\widehat{\mathbf{v}}-\mathbf{v}^*\|_\infty &= \|(\widehat{\mathbf{v}}^T\Sigma_0-\mathbf{e})\Sigma_0^{-1}\|_\infty\\ &\leq \|\Sigma_0^{-1}\|_1(\|\widehat{\mathbf{v}}^TS_0-\mathbf{e}\|_\infty+\|\widehat{\mathbf{v}}\|_1\|S_0-\Sigma_0\|_{\max})\\ &\leq \|\Sigma_0^{-1}\|_1 2\lambda'.\end{aligned}$$

Combining the last bound with (C.5.1), we get:

$$\|\widehat{\mathbf{v}} - \mathbf{v}^*\|_1 \leq 4s_{\mathbf{v}}\|\Sigma_0^{-1}\|_1\lambda',$$

which is what we wanted to show. □

C.6 Proofs for the Quasi-Likelihood

Proof of Theorem 4.8.2. We verify the conditions from Theorem 4.3.3. Under the conditions of Lemma C.6.5 we have that $\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_1 = O_p(\lambda s)$ and hence by Remark C.6.7 we have that:

$$\sqrt{n}\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_1 \sup_{\nu\in[0,1]} \left\|\widehat{\mathbf{v}}^T n^{-1}\sum_{i=1}^n \boldsymbol{X}_i^{\otimes 2} f'(\boldsymbol{X}_i^T\boldsymbol{\beta}_\nu) - \mathbf{e}\right\|_\infty = \sqrt{n}O_p(\lambda')O_p(\lambda s) = o_p(1).$$

Moreover, Lemma C.6.6 gives us that $\|\widehat{\mathbf{v}} - \mathbf{v}^*\|_1 = O_p(\lambda' s_{\mathbf{v}})$, and therefore:

$$\sqrt{n}\|\widehat{\mathbf{v}} - \mathbf{v}^*\|_1 \left\|n^{-1}\sum_{i=1}^n (f(\boldsymbol{X}_i^T\boldsymbol{\beta}^*) - Y_i)\boldsymbol{X}_i\right\|_\infty \leq \sqrt{n}O_p(\lambda' s_{\mathbf{v}})O_p(\lambda) = o_p(1),$$

which completes the proof. □

Proof of Corollary 4.8.4. As in the linear case we only need to verify Lyapunov's condition for the CLT. We have $\Delta = \mathbf{v}^{*T}\boldsymbol{\Sigma}_W\mathbf{v}^* \geq \|\mathbf{v}^*\|_2^2\delta$, by our assumption. Hence $|\Delta|^{3/2} \geq \|\mathbf{v}^*\|_2^3\delta^{3/2}$. Thus we need to control:

$$\frac{n^{-3/2}}{\|\mathbf{v}^*\|_2^3}\sum_{i=1}^n \mathbb{E}|\mathbf{v}^{*T}\boldsymbol{X}_i(f(\boldsymbol{X}_i^T\boldsymbol{\beta}^*) - Y_i)|^3.$$

We note that $\mathbb{E}|\mathbf{v}^{*T}\boldsymbol{X}_i(f(\boldsymbol{X}_i^T\boldsymbol{\beta}^*) - Y_i)|^3 \leq K'^3K^3\|\mathbf{v}^*\|_1^3 \leq K'^3K^3 s_{\mathbf{v}}^{3/2}\|\mathbf{v}^*\|_2^3$. This completes the proof provided that $s_{\mathbf{v}}^{3/2}/\sqrt{n} = o(1)$, which we have assumed. □

Proof of Proposition 4.8.5. We start with the consistency of $\widehat{\Delta}_1$. $\widehat{\Delta}_1$ is an asymptotically consistent

estimate for $\mathbf{v}_1^* = \mathbf{v}^{*T}\mathbf{\Sigma}_W\mathbf{v}^*$, since $|\widehat{\mathbf{v}}_1 - \mathbf{v}_1^*| \leq \|\widehat{\mathbf{v}} - \mathbf{v}^*\|_1 = O_p(\lambda' s_{\mathbf{v}}) = o_p(1)$ under the assumptions of Lemma C.6.6 and $\lambda' s_{\mathbf{v}} = o(1)$.

Next, we consider $\widehat{\Delta}_2$. We have:

$$\widehat{\Delta}_2 = \underbrace{\left(\widehat{\mathbf{v}}^T n^{-1} \sum_{i=1}^n \boldsymbol{X}_i^{\otimes 2} f'(\boldsymbol{X}_i^T \widehat{\boldsymbol{\beta}}) - \mathbf{e}\right) \widehat{\mathbf{v}}}_{I_1} + \widehat{\mathbf{v}}_1.$$

Under the assumptions of Lemma C.6.6 we have $\|\widehat{\mathbf{v}}\|_1 \leq \|\mathbf{v}^*\|_1$ with probability at least $1 - 4d^{-1}$, and therefore $|I_1| \leq \lambda' \|\mathbf{v}^*\|_1 = o(1)$ with probability no less than $1 - 4d^{-1}$. On the other hand as we argued for $\widehat{\Delta}_1$, $\widehat{\mathbf{v}}_1$ is consistent for $\mathbf{v}_1^*$ if $\lambda' s_{\mathbf{v}} = o(1)$ and the assumptions of Lemma C.6.6 hold.

Finally we deal with $\widehat{\Delta}_3$. First we show that we can substitute the term $(Y_i - f(\boldsymbol{X}_i^T \widehat{\boldsymbol{\beta}}))^2$ with $(Y_i - f(\boldsymbol{X}_i^T \boldsymbol{\beta}^*))^2$. We have:

$$\begin{aligned}\widehat{\Delta}_3 = &\underbrace{n^{-1} \sum_{i=1}^n (\widehat{\mathbf{v}}^T \boldsymbol{X}_i)^2 (Y_i - f(\boldsymbol{X}_i^T \boldsymbol{\beta}^*))^2}_{I_2} \\ &+ \underbrace{n^{-1} \sum_{i=1}^n (\widehat{\mathbf{v}}^T \boldsymbol{X}_i)^2 (f(\boldsymbol{X}_i^T \boldsymbol{\beta}^*) - f(\boldsymbol{X}_i^T \widehat{\boldsymbol{\beta}}))(2Y_i - f(\boldsymbol{X}_i^T \boldsymbol{\beta}^*) - f(\boldsymbol{X}_i^T \widehat{\boldsymbol{\beta}}))}_{I_3}.\end{aligned}$$

In Lemma C.6.6 we argued that with probability at least $1 - 4d^{-1}$ we have $\|\widehat{\mathbf{v}}\|_1 \leq \|\mathbf{v}^*\|_1$. Denote this event by E. To this end recall that we are assuming $\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_1 \leq 1$, which enables us to use to use the following bound

$$|f(\boldsymbol{X}_i^T \widehat{\boldsymbol{\beta}}) - f(\boldsymbol{X}_i^T \widehat{\boldsymbol{\beta}})| = |f'(\boldsymbol{X}_i^T \widetilde{\boldsymbol{\beta}})||\boldsymbol{X}_i^T (\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)| \leq C|\boldsymbol{X}_i^T (\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)|,$$

since $|\boldsymbol{X}_i^T\widetilde{\boldsymbol{\beta}}| \leq |\boldsymbol{X}_i^T\boldsymbol{\beta}^*| + \|\boldsymbol{X}_i\|_\infty\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_1 \leq 2K$. Next for I_3 on the event E we have:

$$\begin{aligned}
|I_3| &\leq n^{-1}\sum_{i=1}^n \|\mathbf{v}^*\|_1^2 KC|\boldsymbol{X}_i^T(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)|(2|Y_i - f(\boldsymbol{X}_i^T\boldsymbol{\beta}^*)| + |C\boldsymbol{X}_i^T(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)|) \\
&\leq n^{-1}\sum_{i=1}^n \|\mathbf{v}^*\|_1^2\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_1 2K^2K'C + n^{-1}\sum_{i=1}^n \|\mathbf{v}^*\|_1^2 K^3C^2\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_1^2 \\
&= O_p(\lambda s)\|\mathbf{v}^*\|_1^2 = o_p(1),
\end{aligned}$$

the last equation following from Lemma C.6.5, which holds on the event E. Next for I_2 we have:

$$I_2 = \underbrace{n^{-1}\sum_{i=1}^n (\widehat{\mathbf{v}}^T\boldsymbol{X}_i)^2[(Y_i - f(\boldsymbol{X}_i^T\boldsymbol{\beta}^*))^2 - f'(\boldsymbol{X}_i^T\boldsymbol{\beta}^*)]}_{I_{21}} + \underbrace{n^{-1}\sum_{i=1}^n (\widehat{\mathbf{v}}^T\boldsymbol{X}_i)^2 f'(\boldsymbol{X}_i^T\boldsymbol{\beta}^*)}_{I_{22}}.$$

We first deal with I_{21}. Note that $\|\boldsymbol{X}_i^{\otimes 2}\|_\infty|(Y_i - f(\boldsymbol{X}_i^T\boldsymbol{\beta}^*))^2 - f'(\boldsymbol{X}_i^T\boldsymbol{\beta}^*)| \leq K^2(K'^2 + C)$. Thus combining a union bound with Hoeffding's inequality we get:

$$\mathbb{P}\left(\left\|n^{-1}\sum_{i=1}^n \boldsymbol{X}_i^{\otimes 2}[(Y_i - f(\boldsymbol{X}_i^T\boldsymbol{\beta}^*))^2 - f'(\boldsymbol{X}_i^T\boldsymbol{\beta}^*)]\right\|_\infty \geq t\right) \leq 2d^2\exp\left(-\frac{t^2 n}{2(K^2(K'^2 + C))^2}\right).$$

Selecting $t = \sqrt{6}K^2(K'^2 + C)\sqrt{\frac{\log d}{n}}$ we get that with probability at least $1 - 6d^{-1}$:

$$|I_{21}| \leq \|\mathbf{v}^*\|_1^2\sqrt{6}K^2(K'^2 + C)\sqrt{\frac{\log d}{n}} = o(1).$$

Next from Lemma C.6.1 with probability at least $1 - 2d^{-1}$ we have that:

$$\left\|\frac{1}{n}\sum_{i=1}^n \boldsymbol{X}_i^{\otimes 2} f'(\boldsymbol{X}_i^T\boldsymbol{\beta}^*) - \boldsymbol{\Sigma}_W\right\|_{\max} \leq \sqrt{6}CK^2\sqrt{\frac{\log d}{n}},$$

and hence with probability at least $1 - 4d^{-1}$ (since $\|\widehat{\mathbf{v}}\|_1 \leq \|\mathbf{v}^*\|_1$ is a sub-event of the event

above):

$$|I_{22}| \leq \underbrace{\|\mathbf{v}^*\|_1^2 \sqrt{6} C K^2 \sqrt{\frac{\log d}{n}}}_{o(1)} + \widehat{\mathbf{v}}^T \boldsymbol{\Sigma}_W \widehat{\mathbf{v}}.$$

Finally $\widehat{\mathbf{v}}^T \boldsymbol{\Sigma}_W \widehat{\mathbf{v}} = \mathbf{v}^{*T} \boldsymbol{\Sigma}_W \mathbf{v}^* + 2(\widehat{\mathbf{v}} - \mathbf{v}^*)^T \boldsymbol{\Sigma}_W \mathbf{v}^* + (\widehat{\mathbf{v}} - \mathbf{v}^*)^T \boldsymbol{\Sigma}_W (\widehat{\mathbf{v}} - \mathbf{v}^*)$. Since $|(\widehat{\mathbf{v}} - \mathbf{v}^*)^T \boldsymbol{\Sigma}_W \mathbf{v}^*| \leq \|\widehat{\mathbf{v}} - \mathbf{v}^*\|_1 \|\boldsymbol{\Sigma}_W \mathbf{v}^*\|_\infty = O_p(\lambda' s_{\mathbf{v}}) 1$ by Lemma C.6.6, and $(\widehat{\mathbf{v}} - \mathbf{v}^*)^T \boldsymbol{\Sigma}_W (\widehat{\mathbf{v}} - \mathbf{v}^*) \leq \|\widehat{\mathbf{v}} - \mathbf{v}^*\|_1^2 \|\boldsymbol{\Sigma}_W\|_{\max} = O_p((\lambda' s_{\mathbf{v}})^2) = o_p(1)$. In the last we used that $\|\boldsymbol{\Sigma}\|_W \leq K^2 C = O(1)$. This completes the proof. □

Proof of Corollary 4.8.8. Note that we have shown all required properties of Proposition 4.3.26, but Assumption 4.3.25. By Remark C.6.8 (note that o can be substituted with θ^*) we have $r_5(n) = 2\lambda'$ (defined in Assumption 4.3.25). Thus by Lemma C.6.5:

$$n^{1/2} \lambda' |\widehat{\theta} - \theta^*| \leq n^{1/2} \lambda' \|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_1 \leq n^{1/2} \lambda' \frac{8(2 + CK^3)\lambda s}{\kappa \lambda_{\min}(\boldsymbol{\Sigma}_W)} = o(1),$$

where the last inequality holds with high probability. This concludes the proof. □

Lemma C.6.1. *With probability at least* $1 - 2d^{-1}$ *we have that:*

$$\left\| \frac{1}{n} \sum_{i=1}^n \boldsymbol{X}_i^{\otimes 2} f'(\boldsymbol{X}_i^T \boldsymbol{\beta}^*) - \boldsymbol{\Sigma}_W \right\|_{\max} \leq \sqrt{6} C K^2 \sqrt{\frac{\log d}{n}}.$$

Proof of Lemma C.6.1. By our assumptions we have $\max_i \|\boldsymbol{X}_i^{\otimes 2} f'(\boldsymbol{X}_i^T \boldsymbol{\beta}^*)\|_{\max} \leq CK^2$. Hence by Hoeffding's inequality, and a union bound we have:

$$\mathbb{P}\left(\left\| \frac{1}{n} \sum_{i=1}^n \boldsymbol{X}_i^{\otimes 2} f'(\boldsymbol{X}_i^T \boldsymbol{\beta}^*) - \boldsymbol{\Sigma}_W \right\|_{\max} \geq t \right) \leq 2d^2 \exp\left(-\frac{nt^2}{2C^2 K^4} \right).$$

Setting $t = \sqrt{6} C K^2 \sqrt{\frac{\log d}{n}}$ gives the desired result. □

Lemma C.6.2. *Assume that* $\lambda_{\min}(\mathbf{\Sigma}_W) > 0$ *and* $s\sqrt{\frac{\log d}{n}} \leq (1-\kappa)\frac{\lambda_{\min}(\mathbf{\Sigma}_W)}{(1+\xi)^2\sqrt{6}CK^2}$*, where* $0 < \kappa < 1$*. Then we have that* $\mathrm{RE}_{\mathbf{\Sigma}_{n,W}}(s,\xi) \geq \kappa\,\mathrm{RE}_{\mathbf{\Sigma}_W}(s,\xi) \geq \kappa\lambda_{\min}(\mathbf{\Sigma}_W)$ *with probability at least* $1-2d^{-1}$*.*

Proof of Lemma C.6.2. The proof of this Lemma is identical to the Proof of Lemma C.2.3, so we omit it. □

Definition C.6.3. *Define* $\mathrm{RE}_\kappa(s,\xi) := \kappa\,\mathrm{RE}_{\mathbf{\Sigma}_W}(s,\xi) \geq \kappa\lambda_{\min}(\mathbf{\Sigma}_W)$*.*

Lemma C.6.4. *Let* $\lambda = 2K'K\sqrt{\frac{\log d}{n}}$*. Then with probability at least* $1-2d^{-1}$ *we have:*

$$\|\widehat{\boldsymbol{\beta}}_{S_0^c} - \boldsymbol{\beta}^*_{S_0^c}\|_1 \leq \|\widehat{\boldsymbol{\beta}}_{S_0} - \boldsymbol{\beta}^*_{S_0}\|_1, \tag{C.6.1}$$

and on the same event:

$$\left\| n^{-1}\sum_{i=1}^n (f(\boldsymbol{X}_i^T\boldsymbol{\beta}^*) - f(\boldsymbol{X}_i^T\widehat{\boldsymbol{\beta}}))\boldsymbol{X}_i \right\|_\infty \leq 2\lambda. \tag{C.6.2}$$

Proof of Lemma C.6.4. Note that we have $\max_i |Y_i - f(\boldsymbol{X}_i^T\boldsymbol{\beta}^*)||\boldsymbol{X}_i| \leq K'K$. Hence by Hoeffding's inequality we have:

$$\mathbb{P}\left(\left\| n^{-1}\sum_{i=1}^n (f(\boldsymbol{X}_i^T\boldsymbol{\beta}^*) - Y_i)\boldsymbol{X}_i \right\|_\infty \geq t \right) \leq 2d\exp\left(-\frac{nt^2}{2(K'K)^2}\right).$$

Selecting $t = 2K'K\sqrt{\frac{\log d}{n}}$, yields that with probability at least $1-\frac{2}{d}$ we have $\left\| n^{-1}\sum_{i=1}^n (f(\boldsymbol{X}_i^T\boldsymbol{\beta}^*) - Y_i)\boldsymbol{X}_i \right\|_\infty \leq 2K'K\sqrt{\frac{\log d}{n}}$. Thus when $\lambda = 2K'K\sqrt{\frac{\log d}{n}}$ $\boldsymbol{\beta}^*$ satisfies the Dantzig Selector constraint with probability at least $1-2d^{-1}$. Hence, just as in the proof in Lemma C.2.6, we conclude

(C.6.1) holds with probability at least $1 - 2d^{-1}$. To show (C.6.2) note that:

$$\begin{aligned}\left\| n^{-1} \sum_{i=1}^{n} (f(\boldsymbol{X}_i^T \boldsymbol{\beta}^*) - f(\boldsymbol{X}_i^T \widehat{\boldsymbol{\beta}})) \boldsymbol{X}_i \right\|_\infty &\leq \left\| n^{-1} \sum_{i=1}^{n} (f(\boldsymbol{X}_i^T \widehat{\boldsymbol{\beta}}) - Y_i) \boldsymbol{X}_i \right\|_\infty \\ &\quad + \left\| n^{-1} \sum_{i=1}^{n} (f(\boldsymbol{X}_i^T \boldsymbol{\beta}^*) - Y_i) \boldsymbol{X}_i \right\|_\infty \\ &\leq 2\lambda,\end{aligned}$$

with probability at least $1 - 2d^{-1}$ which completes the proof.

□

Lemma C.6.5. *Suppose that* $s\sqrt{\frac{\log d}{n}} \leq (1-\kappa)\frac{\lambda_{\min}(\boldsymbol{\Sigma}_W)}{(1+\xi)^2\sqrt{6}CK^2}$, $\lambda \leq 1$ *and* $\sqrt{\lambda}s \leq \frac{\mathrm{RE}_\kappa(s,1)}{8(2+CK^3)}$, *where* λ *is selected as in Lemma C.6.4 and* $0 < \kappa < 1$ *is a fixed constant. Then with probability at least* $1 - 4d^{-1}$ *we have:*

$$\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_1 \leq \frac{8(2 + CK^3)\lambda s}{\mathrm{RE}_\kappa(s, 1)}.$$

Proof of Lemma C.6.5. First observe that any point $\widetilde{\boldsymbol{\beta}} = \nu\widehat{\boldsymbol{\beta}} + (1-\nu)\boldsymbol{\beta}^*, 0 \leq \nu \leq 1$, lying on the line segment connecting $\widehat{\boldsymbol{\beta}}$ with $\boldsymbol{\beta}^*$, satisfies the following for all $i = 1, \ldots, n$:

$$0 \leq (\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)^T \boldsymbol{X}_i (f(\boldsymbol{X}_i^T \widetilde{\boldsymbol{\beta}}) - f(\boldsymbol{X}_i^T \boldsymbol{\beta}^*)) \leq (\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)^T \boldsymbol{X}_i (f(\boldsymbol{X}_i^T \widehat{\boldsymbol{\beta}}) - f(\boldsymbol{X}_i^T \boldsymbol{\beta}^*)).$$

This inequality holds since f is increasing, and can be easily verified. This fact combined with (C.6.2) implies that with probability at least $1 - 2d^{-1}$ for any $\widetilde{\boldsymbol{\beta}}$ as above we have:

$$\begin{aligned}0 \leq n^{-1} \sum_{i=1}^{n} (\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)^T \boldsymbol{X}_i (f(\boldsymbol{X}_i^T \widetilde{\boldsymbol{\beta}}) - f(\boldsymbol{X}_i^T \boldsymbol{\beta}^*)) &\leq n^{-1} \sum_{i=1}^{n} (\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)^T \boldsymbol{X}_i (f(\boldsymbol{X}_i^T \widehat{\boldsymbol{\beta}}) - f(\boldsymbol{X}_i^T \boldsymbol{\beta}^*)) \\ &\leq \|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_1 2\lambda.\end{aligned}$$

Take $\nu = \frac{\tau}{\tau + \|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_1}$ with $\tau \leq 1$ in the definition of $\widetilde{\boldsymbol{\beta}}$, and note that in this case we have $\|\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_1 = \nu\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_1 \leq \tau$. Next by the mean value theorem we have:

$$n^{-1}\sum_{i=1}^{n}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)^T\boldsymbol{X}_i(f(\boldsymbol{X}_i^T\widetilde{\boldsymbol{\beta}}) - f(\boldsymbol{X}_i^T\boldsymbol{\beta}^*)) = \underbrace{n^{-1}\nu\sum_{i=1}^{n}((\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)^T\boldsymbol{X}_i)^2 f'(\boldsymbol{X}_i^T\widetilde{\widetilde{\boldsymbol{\beta}}})}_{I},$$

where $\widetilde{\widetilde{\boldsymbol{\beta}}}$ is a point on the line segment between $\widetilde{\boldsymbol{\beta}}$ and $\boldsymbol{\beta}^*$, and thus $\|\widetilde{\widetilde{\boldsymbol{\beta}}} - \boldsymbol{\beta}^*\|_1 \leq \|\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_1 \leq \tau$. We proceed bounding the expression from below:

$$\begin{aligned} I = &\underbrace{\nu(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)^T\boldsymbol{\Sigma}_{n,W}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)}_{I_1} \\ &\underbrace{+ n^{-1}\nu\sum_{i=1}^{n}((\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)^T\boldsymbol{X}_i)^2(f'(\boldsymbol{X}_i^T\widetilde{\widetilde{\boldsymbol{\beta}}}) - f'(\boldsymbol{X}_i^T\boldsymbol{\beta}^*))}_{I_2}. \end{aligned}$$

By Lemma C.6.2 we have that $I_1 \geq \nu\,\mathrm{RE}_\kappa(s,1)\|\widehat{\boldsymbol{\beta}}_{S_0} - \boldsymbol{\beta}^*_{S_0}\|_2^2$ with probability at least $1 - 2d^{-1}$. For I_2, first note that for any $\breve{\boldsymbol{\beta}}$ on the line segment between $\widetilde{\widetilde{\boldsymbol{\beta}}}$ and $\boldsymbol{\beta}^*$ we have:

$$|\boldsymbol{X}_i^T\breve{\boldsymbol{\beta}}| \leq |\boldsymbol{X}_i^T\boldsymbol{\beta}^*| + \|\boldsymbol{X}_i\|_\infty\|\widetilde{\widetilde{\boldsymbol{\beta}}} - \boldsymbol{\beta}^*\|_1 \leq (1+\tau)K \leq 2K.$$

Hence by the Lipschitz property of f' we have:

$$I_2 \geq -Cn^{-1}\nu^2\sum_{i=1}^{n}|(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)^T\boldsymbol{X}_i|^3 \geq -C\nu^2K^3\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|_1^3 \geq -CK^3\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_1\tau^2.$$

Combining the inequalities above, by a union bound we get with probability at least $1 - 4d^{-1}$:

$$\nu\,\mathrm{RE}_\kappa(s,1)\|\widehat{\boldsymbol{\beta}}_{S_0} - \boldsymbol{\beta}^*_{S_0}\|_2^2 \leq CK^3\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_1\tau^2 + \|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_1 2\lambda.$$

Selecting $\tau = \sqrt{\lambda}$ we obtain:

$$\begin{aligned}\mathrm{RE}_\kappa(s,1)\|\widehat{\boldsymbol{\beta}}_{S_0} - \boldsymbol{\beta}^*_{S_0}\|_2^2 &\leq (2+CK^3)\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_1\sqrt{\lambda}(\sqrt{\lambda} + \|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_1) \\ &\leq 2(2+CK^3)\lambda\sqrt{s}\|\widehat{\boldsymbol{\beta}}_{S_0} - \boldsymbol{\beta}^*_{S_0}\|_2 + 4(2+CK^3)\sqrt{\lambda}s\|\widehat{\boldsymbol{\beta}}_{S_0} - \boldsymbol{\beta}^*_{S_0}\|_2^2,\end{aligned}$$

where we used that on the intersection event we have $\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_1 \leq 2\|\widehat{\boldsymbol{\beta}}_{S_0} - \boldsymbol{\beta}^*_{S_0}\|_1$ by Lemma C.6.4. Since we are assuming we are in the regime $\sqrt{\lambda}s \leq \frac{\mathrm{RE}_\kappa(s,1)}{8(2+CK^3)}$, we get:

$$\|\widehat{\boldsymbol{\beta}}_{S_0} - \boldsymbol{\beta}^*_{S_0}\|_2 \leq \frac{4(2+CK^3)\lambda\sqrt{s}}{\mathrm{RE}_\kappa(s,1)},$$

which finally implies:

$$\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_1 \leq \frac{8(2+CK^3)\lambda s}{\mathrm{RE}_\kappa(s,1)}.$$

This completes the proof. □

Lemma C.6.6. *Assume the same assumptions as in Lemma C.6.5 and that* $\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_1 \leq 1$. *Let* $\lambda' \geq \|\mathbf{v}^*\|_1\left(\frac{8(2+CK^3)\lambda s}{\mathrm{RE}_\kappa(s,1)} + \sqrt{6}CK^2\sqrt{\frac{\log d}{n}}\right)$. *Then we have:*

$$\|\widehat{\mathbf{v}} - \mathbf{v}^*\|_1 \leq \frac{8\lambda' s_{\mathbf{v}}}{\mathrm{RE}_\kappa(s,1)},$$

with probability at least $1 - 4d^{-1}$.

Proof of Lemma C.6.6. We start by showing that the $\mathbf{v}^*$ satisfies the constraint with high probability. Note that, by Lemma C.6.1, with probability at least $1 - 2d^{-1}$ we have:

$$\left\|\mathbf{v}^{*T} n^{-1}\sum_{i=1}^n \boldsymbol{X}_i^{\otimes 2} f'(\boldsymbol{X}_i^T\boldsymbol{\beta}^*) - \mathbf{e}\right\|_\infty \leq \|\mathbf{v}^*\|_1\left\|n^{-1}\sum_{i=1}^n \boldsymbol{X}_i^{\otimes 2} f'(\boldsymbol{X}_i^T\boldsymbol{\beta}^*) - \boldsymbol{\Sigma}_W\right\|_{\max} \leq \|\mathbf{v}^*\|_1\sqrt{6}CK^2\sqrt{\frac{\log d}{n}}.$$

Furthermore by the boundedness and Lipschitz assumptions and Lemma C.6.5, we have that the

following bound holds with probability no smaller than $1 - 4d^{-1}$:

$$\left\| n^{-1} \sum_{i=1}^{n} \boldsymbol{X}_i^{\otimes 2} [f'(\boldsymbol{X}_i^T \boldsymbol{\beta}^*) - f'(\boldsymbol{X}_i^T \widehat{\boldsymbol{\beta}})] \right\|_{\max} \leq CK^3 \|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_1 \leq \frac{8CK^3(2 + CK^3)\lambda s}{\mathrm{RE}_\kappa(s, 1)}. \tag{C.6.3}$$

Combining the last two inequalities with a triangle inequality gives:

$$\left\| \mathbf{v}^{*T} n^{-1} \sum_{i=1}^{n} \boldsymbol{X}_i^{\otimes 2} f'(\boldsymbol{X}_i^T \widehat{\boldsymbol{\beta}}) - \mathbf{e} \right\|_\infty \leq \|\mathbf{v}^*\|_1 \left(\frac{8(2 + CK^3)\lambda s}{\mathrm{RE}_\kappa(s, 1)} + \sqrt{6} CK^2 \sqrt{\frac{\log d}{n}} \right),$$

and thus by the assumption on λ' we have that $\mathbf{v}^*$ satisfies the constraint. Hence on the intersection event, i.e. with probability at least $1 - 4d^{-1}$, we have $\|\widehat{\mathbf{v}}\|_1 \leq \|\mathbf{v}^*\|_1$, and similarly to (C.2.6), we have:

$$\|\widehat{\mathbf{v}}_{S_\mathbf{v}^c} - \mathbf{v}^*_{S_\mathbf{v}^c}\|_1 \leq \|\widehat{\mathbf{v}}_{S_\mathbf{v}} - \mathbf{v}^*_{S_\mathbf{v}}\|_1.$$

Next we deal with the following expression:

$$\begin{aligned} \Bigg\| \underbrace{(\widehat{\mathbf{v}} - \mathbf{v}^*)^T \sum_{i=1}^{n} \boldsymbol{X}_i^{\otimes 2} f'(\boldsymbol{X}_i^T \boldsymbol{\beta}^*)}_{I} \Bigg\|_\infty \leq & \underbrace{\left\| \widehat{\mathbf{v}}^T n^{-1} \sum_{i=1}^{n} \boldsymbol{X}_i^{\otimes 2} f'(\boldsymbol{X}_i^T \boldsymbol{\beta}^*) - \mathbf{e} \right\|_\infty}_{I_1} \\ & + \underbrace{\left\| \mathbf{v}^{*T} n^{-1} \sum_{i=1}^{n} \boldsymbol{X}_i^{\otimes 2} f'(\boldsymbol{X}_i^T \boldsymbol{\beta}^*) - \mathbf{e} \right\|_\infty}_{I_2}. \end{aligned}$$

For I_1, by the triangle inequality we have:

$$\begin{aligned} I_1 &\leq \left\| \widehat{\mathbf{v}}^T n^{-1} \sum_{i=1}^{n} \boldsymbol{X}_i^{\otimes 2} [f'(\boldsymbol{X}_i^T \boldsymbol{\beta}^*) - f'(\boldsymbol{X}_i^T \widehat{\boldsymbol{\beta}})] \right\|_\infty + \lambda' \\ &\leq \|\mathbf{v}^*\|_1 \frac{8CK^3(2 + CK^3)\lambda s}{\mathrm{RE}_\kappa(s, 1)} + \lambda', \end{aligned}$$

the last inequality holding with probability at least $1 - 4d^{-1}$ from (C.6.3) and the fact that $\|\widehat{\mathbf{v}}\|_1 \leq \|\mathbf{v}^*\|_1$. Furthermore on the same event $I_2 \leq \|\mathbf{v}^*\|_1 \sqrt{6} C K^2 \sqrt{\frac{\log d}{n}}$. Adding up the previous inequalities we get that with probability at least $1 - 4d^{-1}$: $\|I\|_\infty \leq 2\lambda'$. We next control $|I(\widehat{\mathbf{v}} - \mathbf{v}^*)|$. By Lemma C.6.2 and what we just concluded with probability no less than $1 - 4d^{-1}$ we have:

$$\mathrm{RE}_\kappa(s,1)\|\widehat{\mathbf{v}}_{S_\mathbf{v}} - \mathbf{v}^*_{S_\mathbf{v}}\|_2^2 \leq |I(\widehat{\mathbf{v}} - \mathbf{v}^*)| \leq \|\widehat{\mathbf{v}} - \mathbf{v}^*\|_1 2\lambda' \leq \sqrt{s_\mathbf{v}}\|\widehat{\mathbf{v}}_{S_\mathbf{v}} - \mathbf{v}^*_{S_\mathbf{v}}\|_2 4\lambda'.$$

A simple calculation finishes the proof. □

Remark C.6.7. *Let $\widetilde{\boldsymbol{\beta}}_\nu = \nu\widehat{\boldsymbol{\beta}}_0 + (1-\nu)\boldsymbol{\beta}^*$, where $\widehat{\boldsymbol{\beta}}_0 = (0, \widehat{\boldsymbol{\gamma}}^T)^T$ and hence $\|\widetilde{\boldsymbol{\beta}}_\nu - \boldsymbol{\beta}^*\|_1 \leq \|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_1$ when $\theta = 0$. Note that the same approach we handled I_1 in the proof of Lemma C.6.6, can be used to show:*

$$\sup_{\nu\in[0,1]} \left\| \widehat{\mathbf{v}}^T n^{-1} \sum_{i=1}^n \boldsymbol{X}_i^{\otimes 2} f'(\boldsymbol{X}_i^T \boldsymbol{\beta}_\nu) - \mathbf{e} \right\|_\infty \leq 2\lambda',$$

with probability not smaller than $1 - 4d^{-1}$.

Remark C.6.8. *In addition to the comment in Remark C.6.7 note that when $\widetilde{\boldsymbol{\beta}}_\nu = \nu\widehat{\boldsymbol{\beta}} + (1-\nu)\widehat{\boldsymbol{\beta}}_0$ we have the same conclusion as $\|\widetilde{\boldsymbol{\beta}}_\nu - \boldsymbol{\beta}^*\|_1 \leq \|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_1$ still holds true when $\theta = 0$.*

References

[1] Abbott, C. A., Malik, R. A., van Ross, E. R., Kulkarni, J., and Boulton, A. J. (2011). Prevalence and characteristics of painful diabetic neuropathy in a large community-based diabetic population in the uk. *Diabetes care*, 34(10):2220–2224.

[2] Alquier, P. and Biau, G. (2013). Sparse single-index model. *The Journal of Machine Learning Research*, 14(1):243–280.

[3] Amini, A. A. and Wainwright, M. J. (2008). High-dimensional analysis of semidefinite relaxations for sparse principal components. In *Information Theory, 2008. ISIT 2008. IEEE International Symposium on*, pages 2454–2458. IEEE.

[4] Ananth, C. V. and Kleinbaum, D. G. (1997). Regression models for ordinal responses: a review of methods and applications. *International journal of epidemiology*, 26(6):1323–1333.

[5] Anderson, T. W. (1958). An introduction to multivariate statistical analysis.

[6] Bartlett, P. L., Jordan, M. I., and McAuliffe, J. D. (2006). Convexity, classification, and risk bounds. *Journal of the American Statistical Association*, 101(473):138–156.

[7] Belloni, A., Chen, D., Chernozhukov, V., and Hansen, C. (2012). Sparse models and methods for optimal instruments with an application to eminent domain. *Econometrica*, 80(6):2369–2429.

[8] Belloni, A., Chernozhukov, V., and Wei, Y. (2013). Honest confidence regions for logistic regression with a large number of controls. *arXiv preprint arXiv:1304.3969*.

[9] Bickel, P. J., Ritov, Y., and Tsybakov, A. B. (2009). Simultaneous analysis of lasso and dantzig selector. *The Annals of Statistics*, pages 1705–1732.

[10] Bradic, J., Fan, J., and Wang, W. (2011). Penalized composite quasi-likelihood for ultrahigh dimensional variable selection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 73(3):325–349.

[11] Bunea, F., Tsybakov, A., and Wegkamp, M. (2007). Sparsity oracle inequalities for the lasso. *Electronic Journal of Statistics*, 1:169–194.

[12] Cai, T. and Liu, W. (2011). A direct estimation approach to sparse linear discriminant analysis. *Journal of the American Statistical Association*, 106(496):1566–1577.

[13] Cai, T., Liu, W., and Luo, X. (2011). A constrained l1 minimization approach to sparse precision matrix estimation. *Journal of the American Statistical Association*, 106(494):594–607.

[14] Candes, E. and Tao, T. (2007). The dantzig selector: Statistical estimation when p is much larger than n. *The Annals of Statistics*, pages 2313–2351.

[15] Collobert, R., Sinz, F., Weston, J., and Bottou, L. (2006). Trading convexity for scalability. In *Proceedings of the 23rd international conference on Machine learning*, pages 201–208. ACM.

[16] Cover, T. M. and Thomas, J. A. (2012). *Elements of information theory*. John Wiley & Sons.

[17] d'Aspremont, A., Bach, F., and Ghaoui, L. E. (2008). Optimal solutions for sparse principal component analysis. *The Journal of Machine Learning Research*, 9:1269–1294.

[18] d'Aspremont, A., El Ghaoui, L., Jordan, M. I., and Lanckriet, G. R. (2007). A direct formulation for sparse pca using semidefinite programming. *SIAM review*, 49(3):434–448.

[19] Deshpande, Y. and Montanari, A. (2013). Sparse pca via covariance thresholding. *arXiv preprint arXiv:1311.5179*.

[20] Fan, J. and Lv, J. (2011). Nonconcave penalized likelihood with np-dimensionality. *Information Theory, IEEE Transactions on*, 57(8):5467–5484.

[21] Fang, K.-T., Kotz, S., and Ng, K. W. (1990). *Symmetric multivariate and related distributions*. Chapman and Hall.

[22] Freund, Y. and Schapire, R. E. (1995). A desicion-theoretic generalization of on-line learning and an application to boosting. In *Computational learning theory*, pages 23–37. Springer.

[23] Friedman, J., Hastie, T., and Tibshirani, R. (2008). Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, 9(3):432–441.

[24] Friedman, J. H. (2001). Greedy function approximation: a gradient boosting machine. *Annals of Statistics*, pages 1189–1232.

[25] Gai, Y., Zhu, L., and Lin, L. (2013). Model selection consistency of dantzig selector. *Statistica Sinica*, 23:615–634.

[26] Galer, B. S., Gianas, A., and Jensen, M. P. (2000). Painful diabetic polyneuropathy: epidemiology, pain description, and quality of life. *Diabetes research and clinical practice*, 47(2):123–128.

[27] Godambe, V. P. and Heyde, C. C. (2010). Quasi-likelihood and optimal estimation. In *Selected Works of CC Heyde*, pages 386–399. Springer.

[28] Hall, P. and Heyde, C. C. (1980). *Martingale limit theory and its application*. Academic press New York.

[29] Han, F., Lu, H., and Liu, H. (2014). A direct estimation of high dimensional a direct estimation of high dimensional stationary vector autoregressions. *arXiv preprint arXiv:1307.0293v3*.

[30] Hastie, T., Tibshirani, R., and Friedman, J. (2009). *The Elements of Statistical Learning*. Springer, New York.

[31] Hoeffding, W. (1948). A class of statistics with asymptotically normal distribution. *The Annals of Mathematical Statistics*, pages 293–325.

[32] Hoeffding, W. (1963). Probability inequalities for sums of bounded random variables. *Journal of the American statistical association*, 58(301):13–30.

[33] Hsing, T. and Carroll, R. J. (1992). An asymptotic theory for sliced inverse regression. *The Annals of Statistics*, pages 1040–1061.

[34] Hsu, C.-W. and Lin, C.-J. (2002). A comparison of methods for multiclass support vector machines. *Neural Networks, IEEE Transactions on*, 13(2):415–425.

[35] Jankova, J. and van de Geer, S. (2014). Confidence intervals for high-dimensional inverse covariance estimation. *arXiv preprint arXiv:1403.6752*.

[36] Javanmard, A. and Montanari, A. (2013). Confidence intervals and hypothesis testing for high-dimensional regression. *arXiv preprint arXiv:1306.3171*.

[37] Johnstone, I. M. and Lu, A. Y. (2004). Sparse principal components analysis. *Unpublished manuscript*.

[38] Johnstone, I. M. and Lu, A. Y. (2009). On consistency and sparsity for principal components analysis in high dimensions. *Journal of the American Statistical Association*, 104(486).

[39] Knight, K. and Fu, W. (2000). Asymptotics for lasso-type estimators. *Annals of Statistics*, pages 1356–1378.

[40] Koltchinskii, V. et al. (2009). The dantzig selector and sparsity oracle inequalities. *Bernoulli*, 15(3):799–828.

[41] Krauthgamer, R., Nadler, B., and Vilenchik, D. (2013). Do semidefinite relaxations really solve sparse pca? *arXiv preprint arXiv:1306.3690*.

[42] Laurent, B. and Massart, P. (2000). Adaptive estimation of a quadratic functional by model selection. *Annals of Statistics*, pages 1302–1338.

[43] Ledoux, M. (2005). *The concentration of measure phenomenon*. Number 89 in Mathematical Surveys and Monographs. American Mathematical Society.

[44] Lee, J. D., Sun, D. L., Sun, Y., and Taylor, J. E. (2013). Exact inference after model selection via the lasso. *arXiv preprint arXiv:1311.6238*.

[45] Lee, Y., Lin, Y., and Wahba, G. (2004). Multicategory support vector machines: Theory and application to the classification of microarray data and satellite radiance data. *Journal of the American Statistical Association*, 99(465):67–81.

[46] Li, K.-C. (1991). Sliced inverse regression for dimension reduction. *Journal of the American Statistical Association*, 86(414):316–327.

[47] Li, K.-C. and Duan, N. (1989). Regression analysis under link violation. *The Annals of Statistics*, pages 1009–1052.

[48] Li, L. and Nachtsheim, C. J. (2006). Sparse sliced inverse regression. *Technometrics*, 48(4).

[49] Lin, Y. (2002). Support vector machines and the bayes rule in classification. *Data Mining and Knowledge Discovery*, 6:259–275.

[50] Lin, Y. (2004). A note on margin-based loss functions in classification. *Statistics & Probability Letters*, 68(1):73 – 82.

[51] Liu, H., Han, F., Yuan, M., Lafferty, J., Wasserman, L., et al. (2012a). High-dimensional semiparametric gaussian copula graphical models. *The Annals of Statistics*, 40(4):2293–2326.

[52] Liu, H., Han, F., and Zhang, C.-h. (2012b). Transelliptical graphical models. In *Advances in Neural Information Processing Systems*, pages 809–817.

[53] Liu, H., Lafferty, J., and Wasserman, L. (2009). The nonparanormal: Semiparametric estimation of high dimensional undirected graphs. *The Journal of Machine Learning Research*, 10:2295–2328.

[54] Liu, Y. (2007). Fisher consistency of multicategory support vector machines. In *International Conference on Artificial Intelligence and Statistics*, pages 291–298.

[55] Lockhart, R., Taylor, J., Tibshirani, R. J., Tibshirani, R., et al. (2014). A significance test for the lasso. *The Annals of Statistics*, 42(2):413–468.

[56] Loh, P.-L. and Wainwright, M. J. (2013). Regularized m-estimators with nonconvexity: Statistical and algorithmic theory for local optima. *arXiv preprint arXiv:1305.2436*.

[57] Mardia, K. V., Kent, J. T., and Bibby, J. M. (1979). *Multivariate analysis*. Academic press.

[58] Martin, C. L., Albers, J., Herman, W. H., Cleary, P., Waberski, B., Greene, D. A., Stevens, M. J., and Feldman, E. L. (2006). Neuropathy among the diabetes control and complications trial cohort 8 years after trial completion. *Diabetes care*, 29(2):340–344.

[59] Masnadi-Shirazi, H. and Vasconcelos, N. (2008). On the Design of Loss Functions for Classification: theory, robustness to outliers, and SavageBoost. In Koller, D., Schuurmans, D., Bengio, Y., and Bottou, L., editors, *NIPS*, pages 1049–1056. Curran Associates, Inc.

[60] Mason, L., Baxter, J., Bartlett, P., and Frean, M. (1999). Boosting algorithms as gradient descent in function space. NIPS.

[61] Massart, P. (1990). The tight constant in the dvoretzky-kiefer-wolfowitz inequality. *The Annals of Probability*, pages 1269–1283.

[62] Meinshausen, N. and Bühlmann, P. (2006). High-dimensional graphs and variable selection with the lasso. *The Annals of Statistics*, 34(3):1436–1462.

[63] Meinshausen, N. and Bühlmann, P. (2010). Stability selection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 72(4):417–473.

[64] Meinshausen, N., Meier, L., and Bühlmann, P. (2009). P-values for high-dimensional regression. *Journal of the American Statistical Association*, 104(488).

[65] Murphy, S. N., Mendis, M. E., Berkowitz, D. A., Kohane, I., and Chueh, H. C. (2006). Integration of clinical and genetic data in the i2b2 architecture. In *AMIA Annual Symposium Proceedings*, volume 2006, page 1040. American Medical Informatics Association.

[66] Murphy, S. N., Weber, G., Mendis, M., Gainer, V., Chueh, H. C., Churchill, S., and Kohane, I. (2010). Serving the enterprise and beyond with informatics for integrating biology and the bedside (i2b2). *Journal of the American Medical Informatics Association*, 17(2):124–130.

[67] Nesterov, Y. (2004). *Introductory lectures on convex optimization*, volume 87. Springer Science & Business Media.

[68] Newey, W. K. and McFadden, D. (1994). Large sample estimation and hypothesis testing. *Handbook of econometrics*, 4:2111–2245.

[69] Ning, Y. and Liu, H. (2014). Sparc: Optimal estimation and asymptotic inference under semiparametric sparsity. *arXiv preprint arXiv:1412.2295*.

[70] Paul, D. (2007). Asymptotics of sample eigenstructure for a large dimensional spiked covariance model. *Statistica Sinica*, 17(4):1617.

[71] Said, G. (2007). Diabetic neuropathy□a review. *Nature Clinical Practice Neurology*, 3(6):331–340.

[72] Shah, R. D. and Samworth, R. J. (2013). Variable selection with error control: another look at stability selection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 75(1):55–80.

[73] Shen, X., Tseng, G. C., Zhang, X., and Wong, W. H. (2003). On ψ-learning. *Journal of the American Statistical Association*, 98(463):724–734.

[74] Sinisi, S. E., Polley, E. C., Petersen, M. L., Rhee, S.-Y., and van der Laan, M. J. (2007). Super learning: an application to the prediction of hiv-1 drug resistance. *Statistical Applications in Genetics and Molecular Biology*, 6(1):7.

[75] Taylor, J., Lockhart, R., Tibshirani, R. J., and Tibshirani, R. (2014). Post-selection adaptive inference for least angle regression and the lasso. *arXiv preprint arXiv:1401.3889*.

[76] Tewari, A. and Bartlett, P. L. (2007). On the consistency of multiclass classification methods. *The Journal of Machine Learning Research*, 8:1007–1025.

[77] Thomas, P. and Eliasson, S. (1984). Diabetic neuropathy. *Peripheral neuropathy*, 2:1773–1810.

[78] Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 267–288.

[79] Tibshirani, R. et al. (1997). The lasso method for variable selection in the cox model. *Statistics in medicine*, 16(4):385–395.

[80] van de Geer, S., Bühlmann, P., and Ritov, Y. (2014). On asymptotically optimal confidence regions and tests for high-dimensional models. *The Annals of Statistics*, 42(3):1166–1202.

[81] van der Laan, M. J., Polley, E. C., and Hubbard, A. E. (2007). Super learner. *Statistical applications in genetics and molecular biology*, 6(1):1–21.

[82] Van der Vaart, A. and Wellner, J. (1996). *Weak Convergence and Empirical Processes*. Springer, New York.

[83] Van der Vaart, A. W. (2000). *Asymptotic statistics*, volume 3. Cambridge university press.

[84] Vershynin, R. (2010). Introduction to the non-asymptotic analysis of random matrices. *arXiv preprint arXiv:1011.3027*.

[85] Voorman, A., Shojaie, A., and Witten, D. (2014). Inference in high dimensions with the penalized score test. *arXiv preprint arXiv:1401.2678*.

[86] Wainwright, M. J. (2009a). Information-theoretic limits on sparsity recovery in the high-dimensional and noisy setting. *Information Theory, IEEE Transactions on*, 55(12):5728–5741.

[87] Wainwright, M. J. (2009b). Sharp thresholds for high-dimensional and noisy sparsity recovery using li-constrained quadratic programming (lasso). *Information Theory, IEEE Transactions on*, 55(5):2183–2202.

[88] Wang, Z. (2012). Multi-class hingeboost. method and application to the classification of cancer types using gene expression data. *Methods of information in medicine*, 51(2):162–167.

[89] Wasserman, L. and Roeder, K. (2009). High dimensional variable selection. *Annals of statistics*, 37(5A):2178.

[90] Yang, Y. and Barron, A. (1999). Information-theoretic determination of minimax rates of convergence. *Annals of Statistics*, pages 1564–1599.

[91] Ye, F. and Zhang, C.-H. (2010). Rate minimaxity of the lasso and dantzig selector for the lq loss in lr balls. *The Journal of Machine Learning Research*, 9999:3519–3540.

[92] Yu, B. (1997). Assouad, fano, and le cam. In *Festschrift for Lucien Le Cam*, pages 423–435. Springer.

[93] Yu, Z., Zhu, L., Peng, H., and Zhu, L. (2013). Dimension reduction and predictor selection in semiparametric models. *Biometrika*, page ast005.

[94] Yuan, M. (2010). High dimensional inverse covariance matrix estimation via linear programming. *The Journal of Machine Learning Research*, 11:2261–2286.

[95] Yuan, M. and Lin, Y. (2006). Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 68(1):49–67.

[96] Zhang, C.-H. and Zhang, S. S. (2014). Confidence intervals for low dimensional parameters in high dimensional linear models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 76(1):217–242.

[97] Zhang, Y. and Ghaoui, L. E. (2011). Large-scale sparse principal component analysis with application to text data. In *Advances in Neural Information Processing Systems*, pages 532–539.

[98] Zhao, P. and Yu, B. (2006). On model selection consistency of lasso. *The Journal of Machine Learning Research*, 7:2541–2563.

[99] Zhong, W., Zhang, T., Zhu, Y., and Liu, J. S. (2012). Correlation pursuit: forward stepwise variable selection for index models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 74(5):849–870.

[100] Zhu, J., Rosset, S., Zou, H., and Hastie, T. (2009). Multi-class adaboost. *Statistics and Its Interface*, 2:349–360.

[101] Zhu, L., Miao, B., and Peng, H. (2006). On sliced inverse regression with high-dimensional covariates. *Journal of the American Statistical Association*, 101(474).

[102] Ziegler, D., Gries, F., Spüler, M., and Lessmann, F. (1992). The epidemiology of diabetic neuropathy. *Journal of diabetes and its complications*, 6(1):49–57.

[103] Zou, H. and Yuan, M. (2008). Composite quantile regression and the oracle model selection theory. *The Annals of Statistics*, 36(3):1108–1126.

[104] Zou, H., Zhu, J., and Hastie, T. (2008). New multicategory boosting algorithms based on multicategory fisher-consistent losses. *The Annals of Applied Statistics*, pages 1290–1306.